TELECOMMUNICATION NETWORKS AND INFORMATICS SERVICES

1. BASICS

For the understanding of the design, deployment, operation and maintenance of telecommunications networks and information technology services some fundamental knowledge is indispensable. The objective of the present Chapter 1 is to give an overview of that set of knowledge that seems to be the most important, as judged by the Editorial Board and the editor of this Chapter. All the topics touched upon have solid theoretical foundation. Still the title of the Chapter does not include the term ″theoretical″. There are two reasons for that. On one hand, limitations of size do not allow to present the principal theorems and rules of telecommunications networks and information technology services tracing them back to their roots in mathematics, physics and economic services, even if in some places there are short references to those traces. On the other hand, the said principal theorems and rules themselves give direct hints concerning their applications.

Another reason to start this handbook with a chapter that tries to collect the ″basics″ is that the vast field of telecommunications networks and information technology services can hardly be presented in a strictly didactical way thoughout the whole book. But if in the first Chapter the basics are summarized, elements of the basics can be referred to at a given point in the book even if the detailed treatment of that particular element appears in a later stage in the book.

It was also the intention in the selection of topics for this ″Basics″ Chapter to express some conceptual issues. Perhaps the first of these issues is the convergence of telecommunications, computer technology and content treatment (industry?), their ever increasing interrelationship. Another issue is to increase the ″warehouse″ of classical technology oriented knowledge with aspects of economics, on one hand, and that of quality as well as quality assurance, on the other hand. Within the individual sub-chapters of the chapter topics with a higher probability of significance in the future were preferred on the cost of classical topics that seem to loose from their importance of a very high level is the past.

According to what was said above the first six sub-chapters resemble a summary of communication theory. Signal theory, characterization of signals of natural origin, source coding, channel coding to combat erroneous channel,

modulation methods and radio propagation are topics appearing in all summaries of communication theory. However, because of the great importance of economical bandwith management a separate sub-chapter is devoted to compression techniques. The next five sub-chapter is a direct consequence of the convergence phenomenon. They treat traffic theory beyond the Erlang-world, data protection and public availability focusing on the subject of the data, data security methods that are independent from the subject and content of the data and graph theory applicable to both telecommunications as well as computer networks. The last two sub-chapters are the foundation of telecommunication economics and quality of service assurance, resp.

We sincerely hope that the effort of the authors and reviewers will, for a long period of time, help the distinguished Reader to find her/his way in the ˝infrastructure˝ that is most characteristic of the information society.

Géza Gordos dr., Editor of the Chapter

1.1. Theory of signals

János Levendovszky dr., author

György Dallos dr., reviewer

Modern communication technologies are chiefly engaged with transmitting and processing signals. Regardless of sending information containing voice, video and data; or carrying out traffic management functions in a complicated networking architecture, the underlying task is to transform and process signals. Thus, modern communication architectures are best treated in a system-theoretical framework, where signals are transformed and manipulated to achieve the ultimate goal of Quality of Service communication (i.e., the system complies with certain quality requirements). From this viewpoint, information is represented by data structures, which can either be in analog or in digital form (e.g. analog voice or digitized images). Throughout this book, the objective of transforming signals is

1. to find the best representations of signals which are most suitable for high quality communication;
2. to develop optimal transformations which map one representation of signals into another one to guarantee a reliable and economic transmission of information over the communication channel.

Here reliable communication is defined in terms of achieving a small bit error rate or a tolerated level of delay, whereas the term "economic" stresses the importance of optimal bandwidth allocation (or the optimal usage of resources).

Signal processing theory, as a discipline, deals with representing the signals as analog or digital data structures. Furthermore, it develops abstract properties which can capture how optimal is the representation of a signal. Consequently, any thorough study pursued into communication technologies must start with a brief description of signals and some basic introduction to signal processing theory. Our coverage of this rich and interesting subject is brief and limited in scope. We are mainly concerned with presenting some basic definitions, thus the chapter tries to serve rather as a review than a comprehensive introduction. To come to grips with the underlying material, the readers are assumed to have some prior exposure to elementary calculus, linear algebra and probability theory.

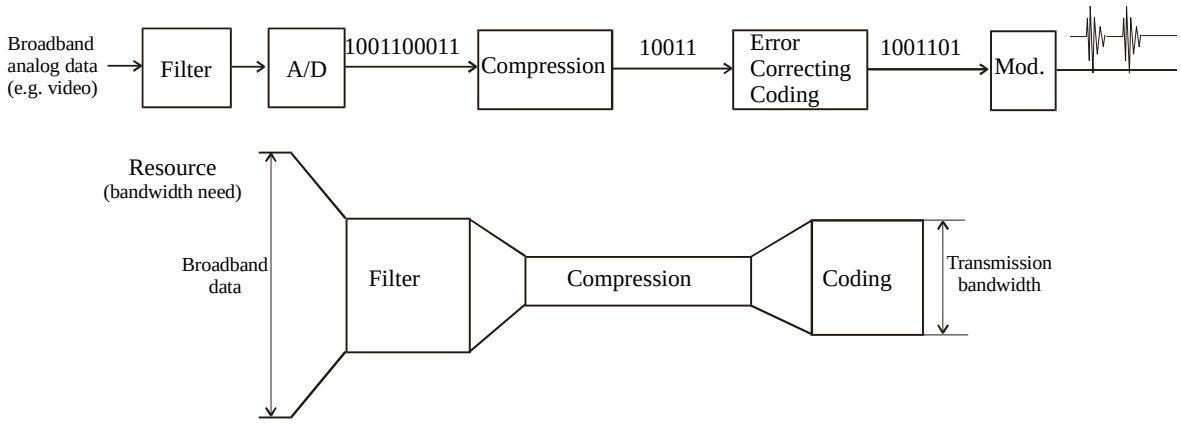


Figure 1.1.1 A typical flow of information and transformation of signals in a digital communication system exemplifying how bandwidth is manipulated to achieve a low rate data transfer

1.1.1. Some fundamental notions of signals and the aim of analysis

In this section, we map out the basic aims of the theory of signals, touching upon the following fields:

- representation of signals;
- architectures for signal processing;
- relevant properties of signals for communication systems (bandwidth, redundancy ...etc.)

Signal processing - the issue of representation

Signals can be regarded as functions of time assuming that information processes (such as voice, video and even data streams) vary in time. (Even the spatial characteristics of images can be turned into time varying signals). Physically, a signal can be a time varying voltage level at the output of a microphone, or a sequence of discrete voltage levels passing through in a digital circuitry. In our viewpoint, however, these phenomena are abstracted into mathematical functions of time, regardless of the physical attributes of the signal. Based on the nature of the input variable (time) and output variable (the value of the function) the following notations are adopted

Communication is chiefly concerned with random signals because information can be interpreted as the amount of randomness in the transmitted messages. Since

Type of signal	Notation
Analog signals	x(t), $0 \le t < \infty$
Sampled analog signals	$x_{k,} k = 0,1,2...$
Digital signals	$\hat{x}_{k,} k = 0,1,2...$
Analog random signals	$\xi(t)\ 0 \le t < \infty$
Sampled random signals	$\xi_k, k = 0,1,2,...$
Digital random signals	$\hat{\xi}_k, k = 0,1,2,...$

in modern communication systems, information is transmitted as binary strings wrapped up in packets, one is faced with the following questions:

- How to convert primarily analog information (like speech, music or video) to digital one?
- What is the amount of information loss (if any) when undergoing such a conversion?
- In what domain signals are to be represented when designing certain elements of a communication systems (e.g. in time domain, frequency domain...etc.)?
- What is the optimal (shortest) representation of signals (e.g., how to make multimedia datastreams suitable for narrowband transmissions) ?

Signal processing - the computational point of view

Signal processing can be basically viewed as a transformation between the input and output signal depicted by Figure 1.1.2. This transformation is always implemented on a computing architecture (e.g., CPU, DSP, an analog filter …etc.)

The next table shows some possible choices of computational architectures for signal processing with the corresponding representations and architectures:

Nature of signal	mathematical representations	Processing algorithms	Architectures	Abstract computing models	Time req. for comp.
analog,	time domain, frequency domain,	analog computation by linear filters	Analog filters	Newton machines	microsec
digital	discrete transform domain (e.g. DFT z transform ...etc)	sequential algorithms written in formal languages Analog connectionist algorithms by PDEs	CPU, DSP CNN chips	Turing machines Trajectory machines	microsec nanosec

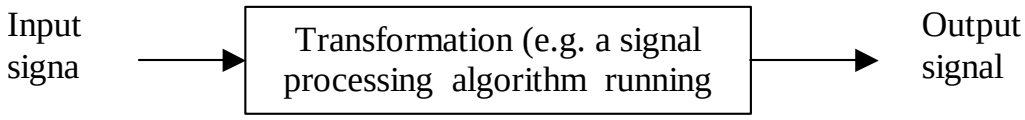


Figure 1.1.2. *Transformation of signals by a given computational architecture*

Based on the table one may conclude that both digital and analog signal processing have their significance in modern communication systems. Despite the fact that information is represented in a digital manner and CPU based architectures or DSPs are widely available, the speed of the necessary computations became a major bottleneck in modern communications. Thus novel approaches to signal processing using connectionist models and analog computational flows can provide an attractive alternative when speed is at the stake. In this chapter we only focus on that part signal-theory which is related to traditional digital signal processing. Some pioneering work which has been recently carried out in the field of analogic computing theory can be found in the work of T. Roska and L. Chua [1.1.1, 1.1.7].

Signal processing as mathematical disciplines for engineering parameters

As was mentioned before, the chapter is not intended to indulge into an overcomplicated formal treatment of signals. The bottomline is to yield results which are of engineering importance. The following table shows, how different mathematical techniques are used to reveal some fundamental properties of signals which are of basic importance when designing a communication system.

Mathematical theory	Engineering output
Fourier transform	Analog bandwidth
Mean square theory (spectral density of stationary processes)	Analog bandwidth of a random process
Sampling theory+Quantization	Digital bandwidth and A/D conversion
Ergod theory of random processes	Adaptation, learning and prediction
Central limit theorem and Gaussian processes	Modeling noise and multiplexed information processes

In order to get a basic understanding of signals first we have to review some elements of deterministic signal processing theory. Then we treat random signals as stochastic processes and describe them by the armory of the so-called "mean-square" theory.

1.1.2. Formal description of analog signals

An analog signal can be represented as a continuous function of time *x(t)*. The basic attributes of *x(t)* are its mean

$$\bar{x} = \lim_{T\to\infty} \frac{1}{T} \int_{-T/2}^{T/2} x(t)dt$$

and average energy

$$\bar{x}^2 = \lim_{T\to\infty} \frac{1}{T} \int_{-T/2}^{T/2} x^2(t)dt.$$

If $x(t)\begin{cases} \neq 0 \text{ if } t \in [T_1, T_2] \\ = 0 \text{ otherwise} \end{cases}$ then interval $[T_1, T_2]$ is referred as the "support" of *x(t)*. Periodical signals are defined by the property of $x(t) = x(t + kT)$, where $k = 0,1,2...$ and they can be expanded into Fourier series, given as follows:

$$x(t) = \sum_k c_k e^{jk2\pi f_0 t} \text{ where } c_k := \frac{1}{T} \int_{-T/2}^{T/2} x(t) e^{-jk2\pi f_0 t} dt \text{ and } f_0 = \frac{1}{T}$$

The transmission capacities of communication channels are often expressed by the notion of bandwidth. Therefore sometimes it is more suitable to represent signals in the frequency domain by using the notion of Fourier transform. If $\int_{-\infty}^{\infty} |x(t)| dt < \infty$, then the following representation exists:

$x(t) = \int_{-\infty}^{\infty} X(f) e^{j2\pi ft} df$, where $X(f) = \int_{-\infty}^{\infty} x(t) e^{-j2\pi ft} dt$.

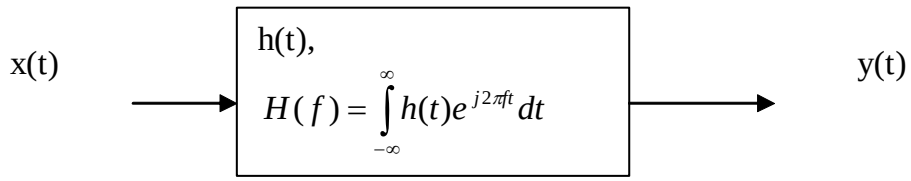
Furthermore, by introducing the complex frequency domain $s = \sigma + j\omega$ one can define the Laplace transform of the so-called switch-on functions ($x(t) = 0$ if $t \le 0$), in the following fashion:

$X(s) = \int_0^{\infty} x(t) e^{-st} dt$. (The strength of the Laplace transform lies in the fact that any function for which $\int_{-\infty}^{\infty} |x(t)e^{-\sigma t}| dt < \infty,\ \sigma \ge 0$ holds has a Laplace transform,

whereas the existence of Fourier transform needed a more stringent condition $\int_{-\infty}^{\infty}|x(t)|dt<\infty$.) The inverse Laplace transform is given as $f(t)=\frac{1}{2\pi j}\int_{c-j\infty}^{c+j\infty}F(s)e^{st}ds$. In spite of the complex integral, there are several techniques have been developed (based on the residuum theorem) how to carry out the inverse transform by elementary algebraic methods. The interested readers are referred to [1.1.5].

Filtering - linear invariant transformations of analog signals

One of the principal methods of analog signal processing is filtering, which is usually meant to shape the frequency spectrum of a signal. If the impulse response function of a filter is denoted by *h(t)* then the linear transformation between the input and output is given by a convolution integral



Figureure 1.1.3 Linear invariant transformation of signals

$y(t)=\int_{-\infty}^{\infty}h(\tau)x(t-\tau)d\tau$ or in frequency domain $Y(f)=H(f)X(f)$. Depending on the nature of $H(f)$ one can carry out low pass, band pass or high pass filtering.

1.1.3. Description and representations of digital signals

As was mentioned above, in modern communication systems information is not transmitted and processed in continuous form. To make benefits of modern digital signal processing algorithms, one has to transform continuous signals into digital domain in two steps (i) sampling and (ii) quantization.

Sampling

The first step in this transformation is sampling, when only signal values taken at discrete times (say $t_k = kT + t_0$) represent the signal. Here t_0 represent the starting point of sampling while *T* is the sampling period. Without loss of generality, we can assume that $t_0 = 0$ and $T = 1$, which means that the analog signal *x(t)* is replaced by its sampled version denoted by x_k.

One must note that in practice we cannot obtain samples described above, as it would need infinitely quick switches. However, samples taken within given period of time - while the switch is closed- results in a sequence $\sum_k x_k h(t-kT)$, where $h(t)$ describes the shape of the samples. The sequence $\sum_k x_k h(t-kT)$ can be viewed as a response of a filter with impulse response function $h(t)$ to the input sequence $\sum_k x_k \delta(t-kT)$. (Where $\delta(t)$ is the well-known Dirac impulse with the property of $\int_{-\infty}^{\infty} \delta(t)dt = 1$). Therefore, another formal definition of a sampled signal is $\tilde{x}(t) = \sum_k x_k \delta(t-kT)$.

One can define the Discrete Fourier Transform of x_k as follows:

$\tilde{X}(f) = T\sum_k x_k e^{j2\pi kfT}$ from which $x_k = \frac{1}{2\pi}\int_0^{1/T} X(f)e^{-j2\pi kfT}df$.

(From the definitions above, it is easy to notice that x_k-s are the Fourier coefficients of $\tilde{X}(f)$ which is a periodic function in the frequency domain with period 1/T). Based on these properties one can derive (see [1.1.5, 1.1.6]) that

$$\tilde{X}(f) = \sum_{k=-\infty}^{\infty} X\left(f + \frac{k}{T}\right).$$

Form this expression one can infer the sampling theorem:

If X(f) is band limited over the frequency interval (-B,B) and $2B \le \frac{1}{T}$, then x(t) can fully be reconstructed from its samples x_k. The frequency $f_0 = \frac{1}{T} = 2B$ is often referred to as the Nyquist rate.

This is due to the fact that sampling over the Nyquist rate ($f_{sampling} > f_0$) implies that

$\tilde{X}(f) = X(f)$ over $f \in (-B, B)$, which means that *X(f)* can be reconstructed from $\tilde{X}(f)$ by a simple low pass filter with cut-off frequency *B*.

It is noteworthy that $\tilde{X}(f) \neq X(f)$ over $f \in (-B, B)$ if $2B > \frac{1}{T}$, meaning the sampling with lower frequency than the Nyquist rate will results in overlapped spectrum from which the original signal cannot be reconstructed.

(In order to solve the problem with an implementable low pass filter one must fulfill strictly $2B < \frac{1}{T}$, where the interval $\frac{1}{T} - B$ is dedicated to accommodate the finite slope of a real low pass filter.)

The important conclusion which emerged that there is no loss of information when sampling an analog signal, if the signal is band limited and the sampling frequency is well beyond Nyquist rate.

Quantization

The next step to transform analog signals into digital ones is by discretizing the amplitude values of the samples. This process is referred to as quantization. In this way, the signal becomes discrete both in time and amplitude and can then be represented by binary strings. This opens up new horizons in signal processing where the full armory of the DSP algorithms can be used. However, "rounding off" the actual signal amplitudes in the course of quantization can lead to unrecoverable distortions imposed on the analog information. This phenomenon is often termed as quantization noise. To analyze this effect in details, let us assume that interval *X* denotes the interval where the sample x_k could fall and it is covered by a discrete grid $\hat{X} = \{C_{-M}, C_{-M+1}, ..., C_{-1}, 0, C_1, ..., C_{M-1}, C_M\}$. Quantization is then defined as

performing the nearest neighbour decision rule on the incoming samples, such as *x* is replaced with C_i where $i = \arg\min_n |x - C_n|$.

(It must be noted that in the forthcoming discussion we drop index *k* from x_k, as quantization is carried out on each sample regardless of the actual time instant).

Quantization error is defined as $\varepsilon := x - C_i$ if sample *x* fell the closest to C_i. Our primary interest is concerned with evaluating the quantization signal-to-noise ratio defined as:

$SNR = \frac{E(\varepsilon^2)}{E(x^2)}$, where *E* stands for the expected value.

This expresses the comparison between the average energy of quantization error an the average signal energy. Here we took into account that if the signal carries information then its sample *x* is a random variable.

The underlying question is how to choose the quantization levels (or $\hat{X}$) in order to lose the least amount of information, or alternatively speaking, how to maximize *SNR*. The set $\hat{X}$ can be regarded as a mapping from uniform quantization levels $\tilde{\hat{X}}$ (where Δ denotes the uniform length quantization intervals), generated by function *f(u).* More precisely, $C_i = f(i\Delta)$, where $\tilde{\hat{X}} = \{-M\Delta, (-M+1)\Delta, ..., -\Delta, 0, \Delta, ..., (M-1)\Delta, M\Delta\}$ corresponds with an equidistant grid in the case of uniform quantization and $\hat{X} = \{f(-M\Delta), f((-M+1)\Delta), ..., f(-\Delta), f(0), f(\Delta), ..., f((M-1)\Delta), f(M\Delta)\}$.

The signal-to-noise ratio associated with a particular *f* is denoted by *SNR(f).* Thus, our strategy is to find:

$$f_{opt} : \max_f SNR(f)$$

$$f_{opt} : \min_f \sum_{i=-M}^{M} E\left((x - C_i)^2 \middle| i = \arg\min_n |x - C_n| \right) p_i$$

Before elaborating on the general solution of the problem, first let us investigate the case of equidistant grid shown below

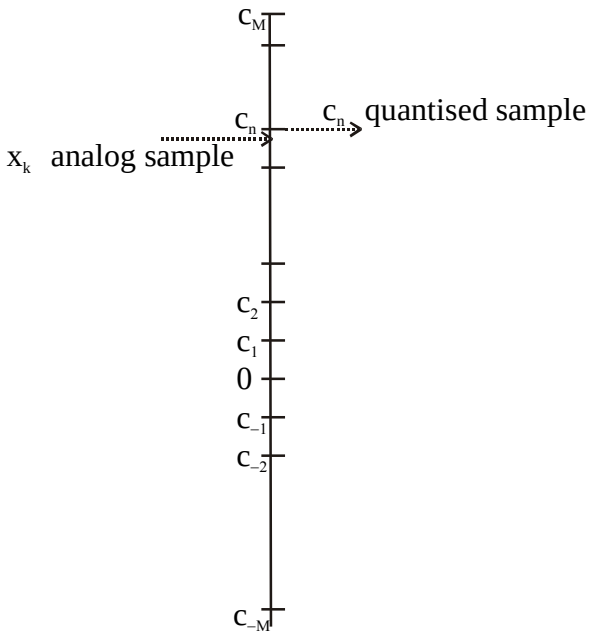


Figure 1.1.4 Uniform quantization of samples

Assuming that ε follows uniform probability distribution

$$E\left(\varepsilon^2\right)=\int_{-\Delta/2}^{\Delta/2} x^2\,\frac{1}{\Delta}dx,$$

whereas the signal energy is $C_M^2/2$ assuming samples from sinusoid signals. This yields $SNR=6(C/\Delta)^2$. If $n:=\log_2(2M+1)$ then one obtains $SNR=\frac{3}{2}2^{2n}$, which describes an important relationship between the number of quantization levels expressed in bits and the quality of quantization (SNR).

Now we can turn our attention to a more general case, namely we want to find the optimal function *y=f(x)* (often called as compression function) which maximizes the signal-to-noise ratio. One can see that if Δ is small enough then $C_i-C_{i-1}=\frac{\Delta}{f^{(-1)}{}'(i\Delta)}=\frac{2}{Nf^{(-1)}{}'(i\Delta)}$ (here $f^{(-1)'}(x)$ denotes the derivative of the inverse function). Thus, the average quantization noise energy in the *i*th interval can be expressed as $E\left(\varepsilon^2\middle|x\in(C_{i-1},C_i\right)=\frac{1}{3N^2\left(\frac{df(x)}{dx}\right)^2}$, which finally yields

$SNR(f) = \dfrac{3N^2 \int_{-1}^{1} p(x)x^2 dx}{\int_{-1}^{1} p(x)\left(\dfrac{df^{-1}(x)}{dx}\right)^2 dx}$. In this way the optimal quantization characteristic can be obtained by solving the following well posed optimization problem:

$f_{opt} : \min_f \dfrac{3N^2 \int_{-1}^{1} p(x)x^2 dx}{\int_{-1}^{1} p(x)\left(\dfrac{df^{-1}(x)}{dx}\right)^2 dx}$. It is still a hard task to determine f_{opt} analytically. However, when the aim is to guarantee the same quantization SNR regardless of the level of the input signal, it is easy to point out that the solution is $f(x) \approx \log x$, which is frequently used in the case of voice communication (see [1.1.5]).

A standard for sampling and quantization of speech signals - PCM

Pulse Code Modulation (PCM) was developed in the seventies for digitized speech and later became the first step towards TDMA hierarchies. The specification of PCM is detailed in the standard of ITU-T G.711.

When implementing PCM, the first step is to filter the analog speech signal in order to suppress the spectral components beyond 4 kHz. This is followed by sampling at the rate 8 kHz, then a uniform 8-bit quantization (the number of quantization levels is $2^8 = 256$). As result, PCM yields 64 Kbps dataspeed. A simple example of digital-to-analog conversion (by using 2- and 3-bit quantizer) is depicted by figure 1.1.5.

Since the uniform quantizer does not provide the best SNR, the quantization is nonlinear by applying a compressor function on the analog samples, which is inverted by an expander in the course of digital-to-analog conversion. (The words „compressor” and „expander” are often combined into the terminology „compander”).

Due to the derivation of the optimal quantizer the compressor has a logarithmic characteristics. The real compressor function in Europe is often called „A-law” and it is given as follows:

$y = \dfrac{1+\log(Ax)}{1+\log(A)}$ if $\dfrac{1}{A} < x \le 1$ or $y = \dfrac{Ax}{1+\log(A)}$ if $0 \le x \le \dfrac{1}{A}$

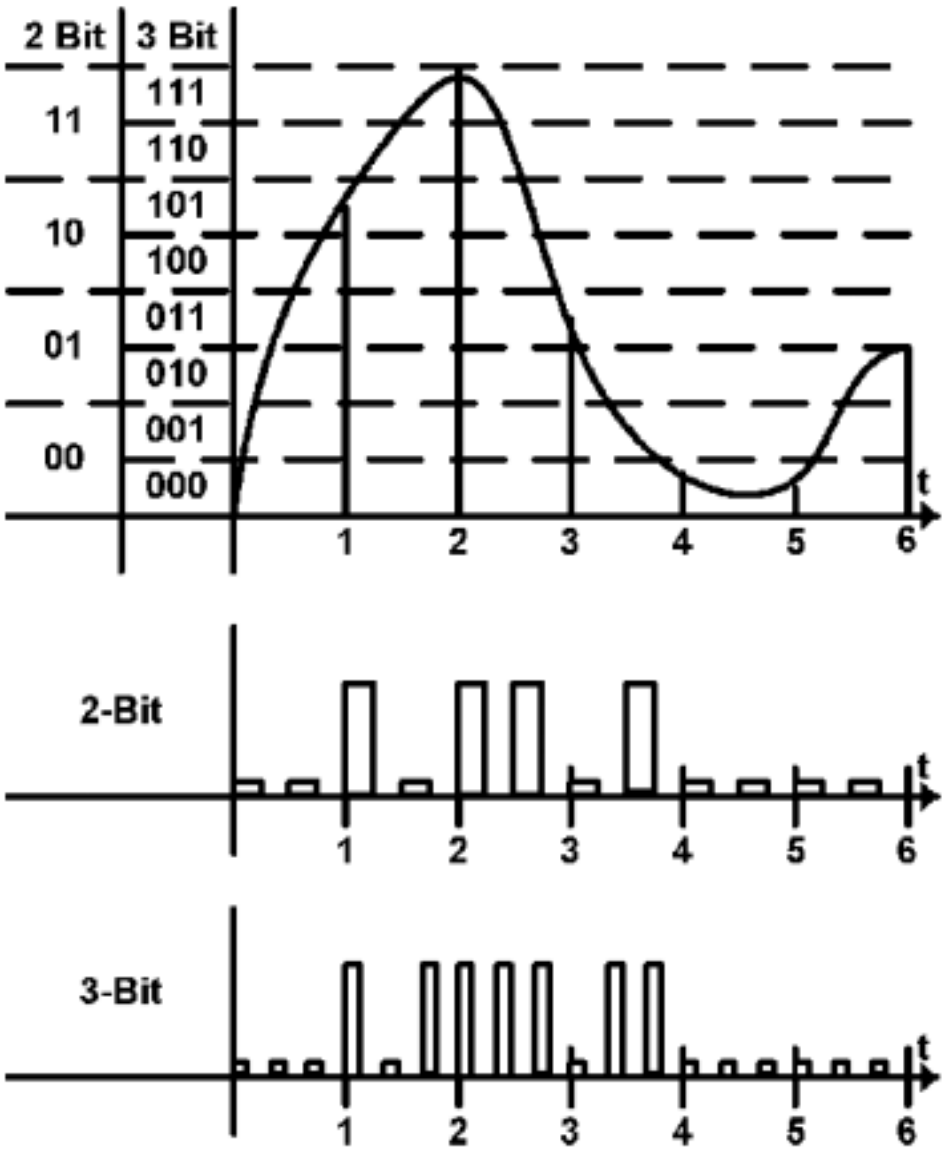


Figure 1.1.5 The representation of an analog signals by digital pulses

In the United States and Japan the so-called „μ-law" came into use, which applies the following formula:

$$y = \frac{\log(1+\mu x)}{\log(1+\mu)}.$$

The specific values are $A = 87.6$ and $\mu = 1000$.

In the course of quantization the signal amplitude range is divided into 16 segments (8 positive and 8 negative segments). Each segment is double size than the previous one and in each segment uniform quantization is performed. The first bit of the 8-bit length corresponding codeword represents the sign of the signal

amplitude, whereas the second, third and fourth identify the segment. The last four bits indicates the quantization level in the corresponding segment.

The „A-law" ensures a wider dynamic range than the „μ-law". However, the „μ-law" yields a better SNR for small amplitude signals than „A-law". In the case of international calls the $A-\mu$ conversion must be introduced (which is the liability of the „μ countries" according to international agreements).

The further refinement of PCM speech coding leads to DPCM, ADPCM and Delta Modulation (for detailed treatment, see Section 1.4.).

1.1.4. Random processes

As was stressed before, information can only be conveyed if a signal exhibits some randomness in its behaviour (to be more precise, the randomness carries the information). Therefore, coming to grips with random processes plays a central role in communication theory. In this section we set out to describe random (or in other word, stochastic) signals. The underlying goal is to arrive at some notions (such as bandwidth, average energy ...etc.) which have technical impact on designing an information transmission system.

Recalling that casting a dice is a random variable yielding a number from the set {1,2,3,4,5,6}, naively one could say, that a stochastic process is a "generalized" random variable which yield not a number but a function as a realization. For instance, the temperature as a function of time during the day is a stochastic process. Even during two consecutive days we can get very different realizations. However, treating stochastic processes as random variables which result in functions yields an almost intractable analysis. Therefore, random processes are better defined in the following way [1.1.3, 1.1.6]:

Definition:

A random process is a set of random variables $\{\xi_t, t \in [0,T]\}$, where at a given point of time $t_1 \in [0.T]$ the random variable ξ_{t_1} expresses the randomness in the value of the signal.

As a result, ξ_{t_1} is often referred to as a "one dimensional projection" of the random process $\{\xi_t, t \in [0,T]\}$.

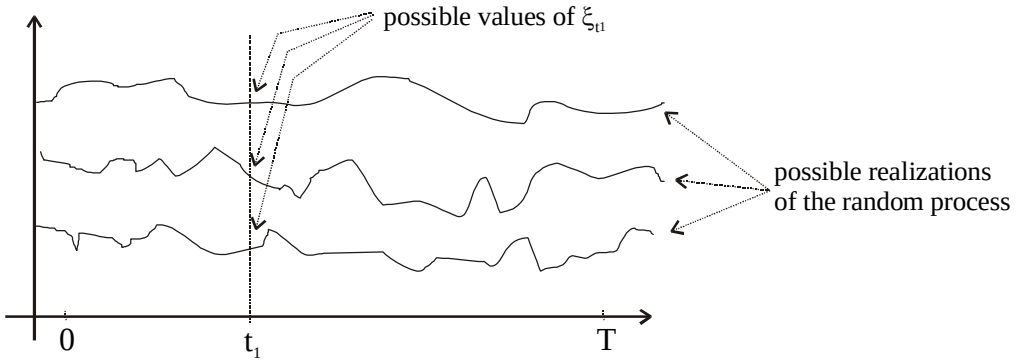


Figure 1.1.6 Realizations of a random process with a one dimensional projection at t_1 as a random variable

Similarly, one can easily introduce multi-dimensional projections, given in the form $(\xi_{t_1}, \xi_{t_2}, ..., \xi_{t_n})$, which denotes the vector valued random variable belonging to the time set $(t_1, t_2, ..., t_n)$. The significance of finite dimensional projections lies in the fact that they are finite dimensional random variables, therefore they can be described by the classical tools of probability theory, such as with their probability distribution functions $F_{t_1,t_2,...,t_n}(x_1, x_2, ..., x_n) = P(\xi_{t_1} < x_1, \xi_{t_2} < x_2, ..., \xi_{t_n} < x_n)$ or with the corresponding probability density function (p.d.f.) $f_{t_1,t_2,...,t_n}(x_1, x_2, ..., x_n) = \dfrac{\partial^n F_{t_1,t_2,...,t_n}(x_1, x_2, ..., x_n)}{\partial x_1 \partial x_2 ... \partial x_n}$.

One can also describe stochastic processes by using first and second order statistics (such as mean and correlation) defined as follows:

$$m(t) := \int_{-\infty}^{\infty} x f_t(x) dx \text{ and } R(t_1, t_2) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x_1 x_2 f_{t_1 t_2}(x_1, x_2) dx_1 dx_2$$

These quantities, of course, do not yield such an in-depth description of the process as the p.d.f.-s of finite dimension projections do.

Treating stochastic processes as ensembles of random variables, one can circumvent the underlying difficulties of introducing probability measures in function spaces (for further details see [1.1.2, 1.1.3]).

Stationarity

One of the fundamental notion of random processes is stationarity. It entails the time invariance of the underlying statistics of a process. For example, if statistical estimation was carried out to ascertain the p.d.f. of the process at a given time instant, how long can we maintain this estimation. In a stationary process what we know statistically about a process at a given time will hold in the future (like casting a dice follows uniform distribution regardless of throwing it on Sunday 5 pm. or on Wednesday 10 am.).

The notion of stationarity depends on what level of statistical descriptors are time invariant. We basically distinguish stationarity at two levels:

- *stationarity in wide sense* when only the expected value and the correlation function is time invariant;
- *stationarity in strict sense* when the p.d.f.s of all finite dimension projection of the process is time invariant.

As a result, a process is stationer in the wide sense if

$m(t) = m(t+\tau) \quad \forall \tau > 0$ and $R(t_1, t_2) = R(t_1 + \tau, t_2 + \tau) \quad \forall \tau > 0$. It is easy to derive that form these conditions the following properties hold

$m(t) = \text{constant}$ and $R(t_1, t_2) = R(t_1 - t_2) = R(\tau)$.

A process is stationer in the strict sense if $F_{t_1,t_2,...,t_n}(x_1, x_2, ..., x_n) = F_{t_1+\tau,t_2+\tau,...,t_n+\tau}(x_1, x_2, ..., x_n) \quad \forall \tau > 0$.

One must note that stationarity in a strict sense implies stationarity in the wide sense (the invariance of expected value and correlation function follows from the invariance of the p.d.f.-s) but the statement in a reverse order may not be necessarily true.

Ergodicity

In order to explore the statistics of a random process, one can observe the process at different time instants and provide empirical statistics based on time averaging. E.g. the expected value of a one dimensional projection is estimated by obeserving the process at ten dieffrnet time instant and then one averages the observed smaples over ten. This gives rise to the fundamental question, to what extent this empirical statistics can be trusted upon? The question is roughly the same

when flipping a single coin a thousand times end determining the empirical average and flipping one thousand identical coins at the same time to determine estimate of the expected value. Is there any difference between the two experiments? The answer lies with ergodicity which claims that for almost any *g* function (to be precise for any *g* Borel measurable function [1.1.2, 1.1.3]) the time averages are equivalent with the statistical averages, given as follows:

$$\lim_{T\to\infty}\frac{1}{T}\int_{-\infty}^{\infty} g\big(x(t)\big)dt = E\big(g(\xi_t)\big)$$

From engineering point of view, the most important implications are when *g* is the identity function implying that

$$\lim_{T\to\infty}\frac{1}{T}\int_{-\infty}^{\infty} x(t)dt = E\big(\xi_t\big)\ ,$$

and when *g* is selected to be quadratic implying that

$$\lim_{T\to\infty}\frac{1}{T}\int_{-\infty}^{\infty} x^2(t)dt = E\big(\xi_t^2\big).$$

This connects the time average to expected value and averaged energy to second order moment. One must also note that

$\lim_{T\to\infty}\frac{1}{T}\int_{-\infty}^{\infty} x^2(t)dt = E\big(\xi_t^2\big) = R(0)$, meaning that the correlation function at zero time lag yields the average signal energy in the case of ergodic processes.

It is easy to see that every ergodic process is stationary in strict sense at the same time and the class of widely stationary processes contains the class of strictly stationary processes. This relationship is shown by the figure 1.1.7.

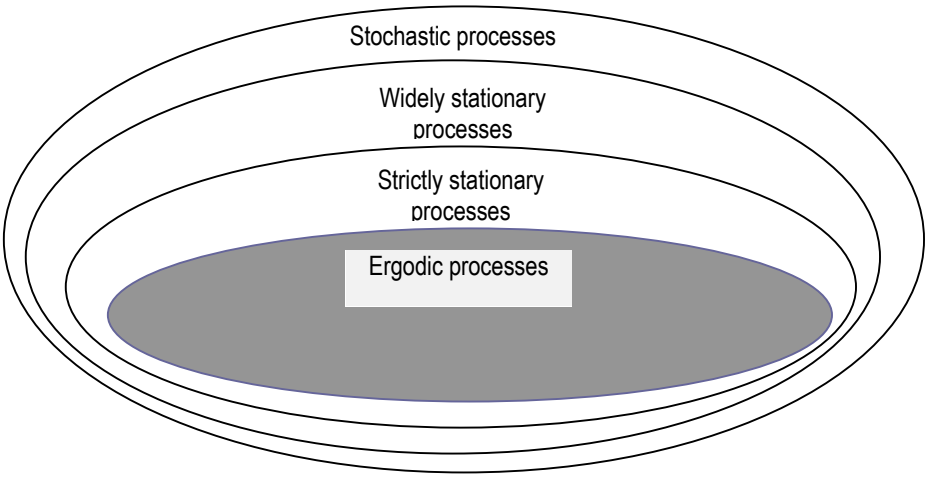


Figure 1.1.7 Classification of random processes

1.1.5. Spectral density and bandwidth of widely stationary random processes

In order to see how the energy distributed in the frequency domain we introduce spectral density as the Fourier transform of the correlation function:

$s(f) = \int_{-\infty}^{\infty} R(\tau) e^{-j2\pi f\tau} d\tau$. This indeed reflects to the spectral distribution energy as

$\lim_{T\to\infty} \frac{1}{T} \int_{-\infty}^{\infty} x^2(t)dt = E\left(\xi_t^2\right) = R(0)$ which can be expressed by inverse Fourier transform as $\lim_{T\to\infty} \frac{1}{T} \int_{-\infty}^{\infty} x^2(t)dt = R(0) = \int_{-\infty}^{\infty} s(f)df$. Therefore there is a mapping form average energy and spectral density. Furthermore, we can define now the bandwidth of a random process as the bandwidth of the spectral density. More precisely, there is no need to transmit an information process on the frequency band where the spectral density is almost zero (there is no component of the average energy on that particular frequency component).

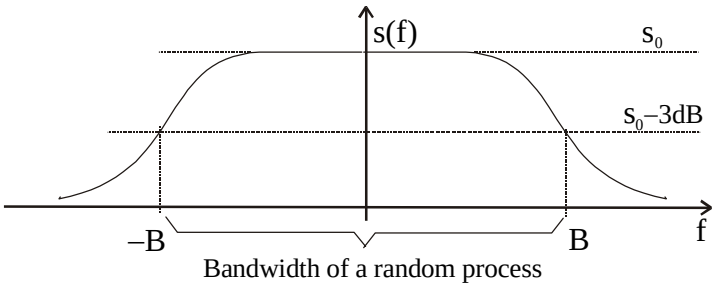


Figure 1.1.8 A band limited random process

1.1.6. Linear invariant transformation of random processes - filtering

Let us assume that random process is transformed by a filter having an impulse response function *h(t)*. Then the relationship between the input and the output is given as

$\eta(t) = \int_{-\infty}^{\infty} h(t-\tau)\xi(\tau)d\tau$. It is easy to prove that the relationship between the output process spectral density and the input process spectral density is given as $s_\eta(f) = |H(f)|^2 s_\xi(f)$, where *H(f)* is the frequency characteristics of the filter defined as $H(f) = \int_{-\infty}^{\infty} h(t)e^{-j2\pi ft}dt$. This rule also helps us to measure the spectral density of an ergodic process as follows:

One can define a narrowband bandpass filter at the frequency f_0 as $H(f) = \begin{cases} 1 \text{if } f_0 - \Delta/2 \le |f| \le f_0 + \Delta/2 \\ 0 \text{ otherwise} \end{cases}$.

Then measuring the average energy at the output of this filter (the output process is denoted by $\eta(t)$ and the corresponding realization is *y(t)*, respectively) , one obtains

$$\lim_{T\to\infty} \frac{1}{T} \int_{-T/2}^{T/2} y^2(t)dt = E(\eta_0^2) = R_\eta(0) = \int_{-\infty}^{\infty} s_\eta(f)df = \int_{-\infty}^{\infty} |H^2(f)| s_\xi(f)df = 2\int_{f_0-\Delta/2}^{f_0+\Delta/2} s_\xi(f)df \approx 2\Delta s_\xi(f_0)$$

This formula means that rigged up with a proper filter the frequency of which f_0 can be adaptively changed and with an "energy-meter", the spectral density of any

ergodic process can be measured. The measuring set up is indicated by the next figure

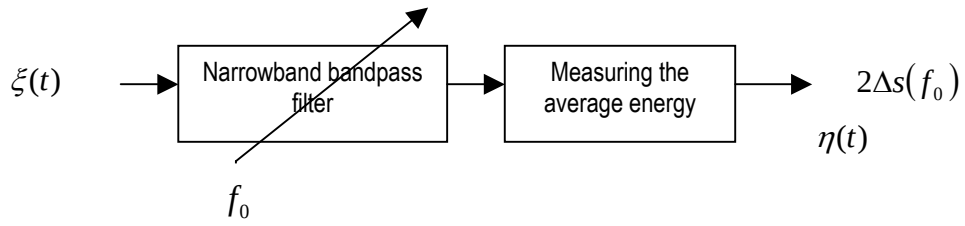


Figure 1.1.9 Measuring the spectral density of a random process

However, with spectral density at hand either the bandwidth or the correlation function can be calculated, which almost "fully" characterizes the process at engineering level. Therefore, the corresponding theory is often labeled as "mean-square theory".

1.1.7. Gaussian processes and white noise

Similarly to the elementary probability theory Gaussian processes play a central role in the theory of random processes. Their importance is twofold:

- a Gaussian process is fully characterized by the mean value function and the correlation function;
- due to the generalization of central limit theorem, the sum of independent processes tend to behave in a Gaussian manner.

A Gaussian process is defined as follows:

For each possible sample set $(t_1,t_2,...,t_n)$ *n*=1,2,.... the vector valued random variable $\bar{\xi} = (\xi_{t_1}, \xi_{t_2},..., \xi_{t_n})$ follows Gaussian distribution with the following p.d.f.

$$f(x_1, x_2,..., x_n) = \frac{1}{\sqrt{(2\pi)^n \det(\mathbf{K})}} e^{-\frac{1}{2}(\mathbf{x}-\mathbf{m})^{\mathrm{T}}\mathbf{K}^{-1}(\mathbf{x}-\mathbf{m})},$$

where **K** is the covariance matrix defined as $K_{ij} = E(\xi_{t_i}\xi_{t_j}) - E(\xi_{t_i})E(\xi_{t_j})$ and vector **m** is the mean vector. This implies that $K_{ij} = R(t_i, t_j) - m(t_i)m(t_j)$. Since the p.d.f.-s of finite dimension projection fully determine the statistical behaviour of the random process and these p.d.f. s can be obtained via the correlation function and expected value function, it is indeed true that a Gaussian process is fully characterized by its mean and correlation function.

Since there is analytical relationship between the p.d.f.-s and mean and correlation, every Gaussian process which is weakly stationer is strongly stationer at the same time (this property does not hold for other stochastic processes). Therefore, in the case of Gaussian processes we simply adopt the terminology "stationary".

We define white Gaussian noise as Gaussian process, the spectral density function of which is constant $s(f) = const \quad f \in (-\infty, \infty)$. As a result the correlation function of a white Gaussian process is the Dirac delta $R(\tau) = const\delta(\tau)$, which entails the whatever samples are chosen, they are uncorrelated (furthermore, since we are talking of Gaussian processes it implies statistical independence). One can see, that white Gaussian noise is rather a useful mathematical abstraction than physical reality as it assumes infinite energy. Nevertheless, the properties of a finite energy, coloured Gaussian noise (which is obtained by filtering a white Gaussian process) can easily be calculated by using the mathematical abstraction of white Gaussian process.

The importance of white Gaussian processes lies in the fact that these processes are good approximation of the thermal noise. These approximation is justified by the central limit theorem (the aggregate effect of randomly fluctuating particle is roughly Gaussian). Furthermore, multiplexing a large number of information processes can also be regarded as a Gaussian process. Therefore, Gaussian processes tend to bear fundamental impact on the statistical design of communication systems.

Summary

In this section the main properties of signals were introduced and analyzed. The emphasis was given to converting analog signals into digital ones. A brief treatise of stochastic processes were also provided with special stress on stationarity and ergodicity.

References

[1.1.1] Chua, L.O. and Roska, T.: "The CNN Paradigm", *IEEE Trans. on Circuits and Systems*, Vol. 40., March, 1993.

[1.1.2] Doob, J. *Stochastic Processes*, New York-London, 1953

[1.1.3] Gihman, I., Szkorohod, A: *Introduction to stochastic processes*, Technical Press, Budapest 1975 (in Hungarian)

[1.1.4] Haykin, S.: *Neural networks - a comprehensive foundation*, Prentice-Hall, 1999

[1.1.5] Proakis, J, Manolakis, D..: *Digital signal processing*, McGraw-Hill, 1996

[1.1.6] Papoulis, A.: *Probability, random variables and stochastic processes*, McGraw-Hill, 1984

[1.1.7] Roska, T. and Chua, L.O.: "The CNN Universal Machine: an Analogic Array Computer", *IEEE Trans. on Circuits and Systems*, Vol. 40., March, 1993.

1.2. Characterizing real-life information processes

János Levendovszky dr., author

György Dallos dr., reviewer

The fundamental goal of designing a communication system is to transmit information in a reliable manner. Reliability in modern communication networking entails that the transmitted information must satisfy certain quality criteria (e.g., a given delay, bit error rate ...etc.). These criteria are often termed as Quality of Service (QoS) requirements. However, in order to achieve a given level of QoS with the smallest amount of resources, one has to carry out a thorough study on real-life information processes, such as *voice, video* and *data.* The aim of the analysis is to reveal the "resource-demand" presented by a given source to the communication network.

"Resource demand" is a rather general terminology including:

- bandwidth;
- signal-to-noise ratio (SNR);
- bit error rate (BER)... etc.

This sort of characterization tends to be even more important in communication standards providing integrated services, like B-ISDN. In this case, the networking architecture which implements integrated services has to handle many different type of sources. Thus, efficient network utilization can only be achieved if proper bandwidth allocation mechanisms are designed for a large variety of sources.

Getting down to modeling, we have to distinguish between the (i) characterization of originally analog sources (voice, music and video) when the "terminal" of the communication line is the real human perception system (the organs of the corresponding perception are the ears and eyes); (ii) and the characterization of data flows (e.g. file transfers, e-mails ..etc.) where information was originally generated in digital form. One must note, however, that in the advent of "convergence" (i.e., telecommunication and broadcasting world is rapidly getting united with data network world), we have to characterize all these sources within the

uniform framework of packet switched communication. In this case the “type” of data is only identified by the corresponding QoS requirements.

In the case of analog sources the primary objective is to transmit information in such a way that it is still "enjoyable" at the receiver side. The degree of being "enjoyable", however, solely hinges upon some subjective criteria which may not lend itself to be expressed in well defined engineering measures (e.g. it is rather difficult to provide a universal measure for the quality of music or speech). On the other hand, engineering design needs exact parameters of "resource-demands”. As a result, the endeavour of characterizing real-life analog processes is to translate the subjective criteria into exact engineering quantities. In order to do this, we have to briefly summarize the characteristics of human perception systems (e.g. ears, eyes) and then by drawing some conclusions we can arrive at the necessary parameters.

To design packet switched networks for a given QoS, the underlying question is the statistics of the source. More precisely, the task is to reveal the statistical parameters of the random flow of data which then determine the bandwidth need and other important QoS measures.

As a consequence, we apply the "layered" modeling approach, a shown in figure 1.2.1.

1.2.1. Audio signals (voice and music)

In this section we characterize voice and music sources with respect to their bandwidth and SNR …etc.

Perceptual characterisation

In order to capture the characteristics of voice and music which are of engineering relevance first we have to elaborate on the perceptual properties of human audition.

The frequency characteristics of ears together with typical frequency range of voice and music are depicted by the figure 1.2.2:

From this figure, one can infer that speech falls in the frequency domain of (20Hz, 3400 Hz) and music is (10Hz, 20 KHz).

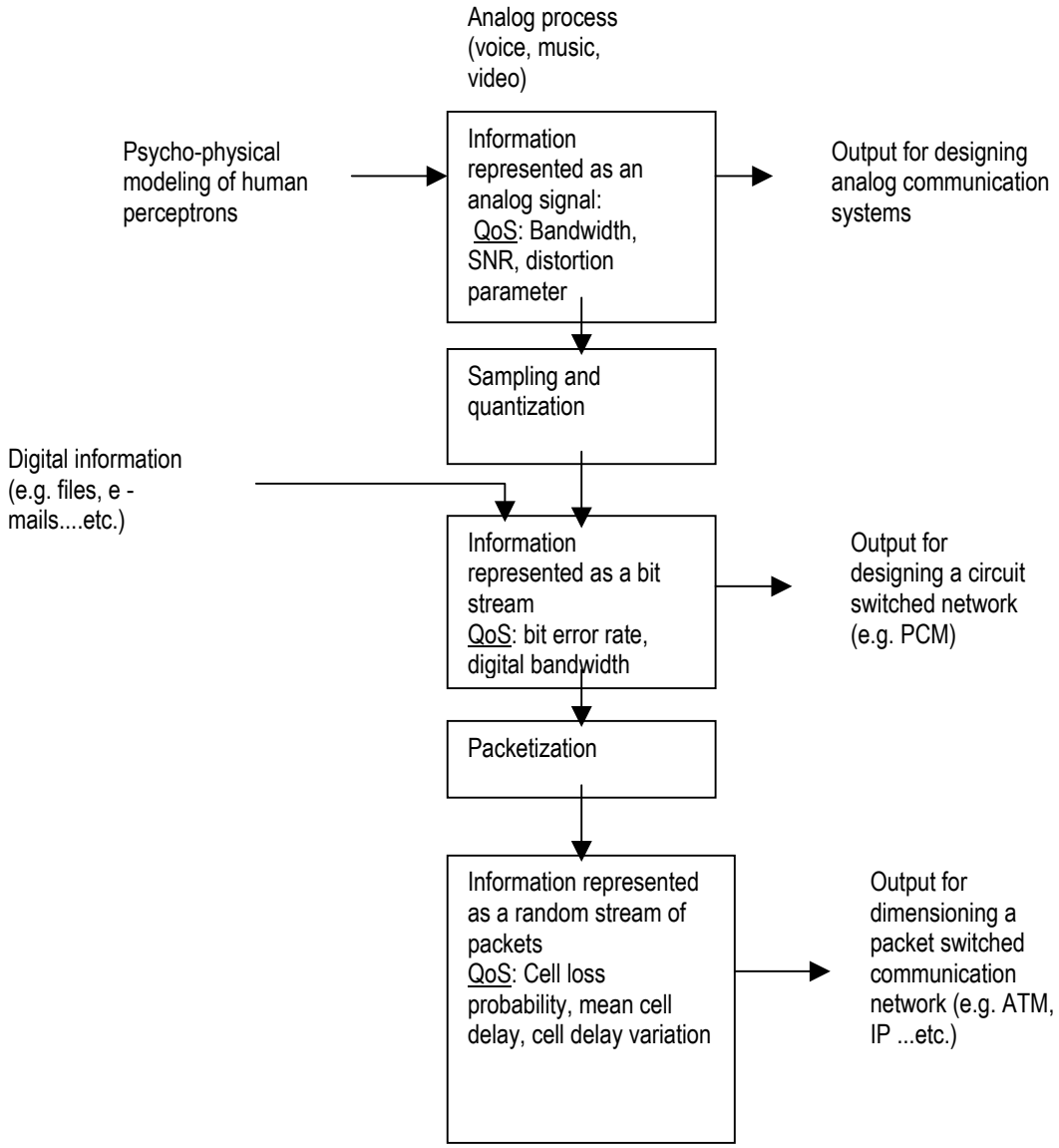


Figure 1.2.1 Level of source modelin

Quality demands and transmission properties of speech

Analog voice is also characterized by the tolerated level of SNR. This usually falls in the range of 20 - 25 dB in the case of telephony.

Another fundamental property which determine the quality of analog voice is the nonlinear distortion parameter. Nonlinear distortion is quantified by investigating the "amount" of higher frequency components at the output given in response to a sinusoidal in the input. In this case, the occurrence of the higher frequency components at the output are solely due to the nonlinear characteristics of the

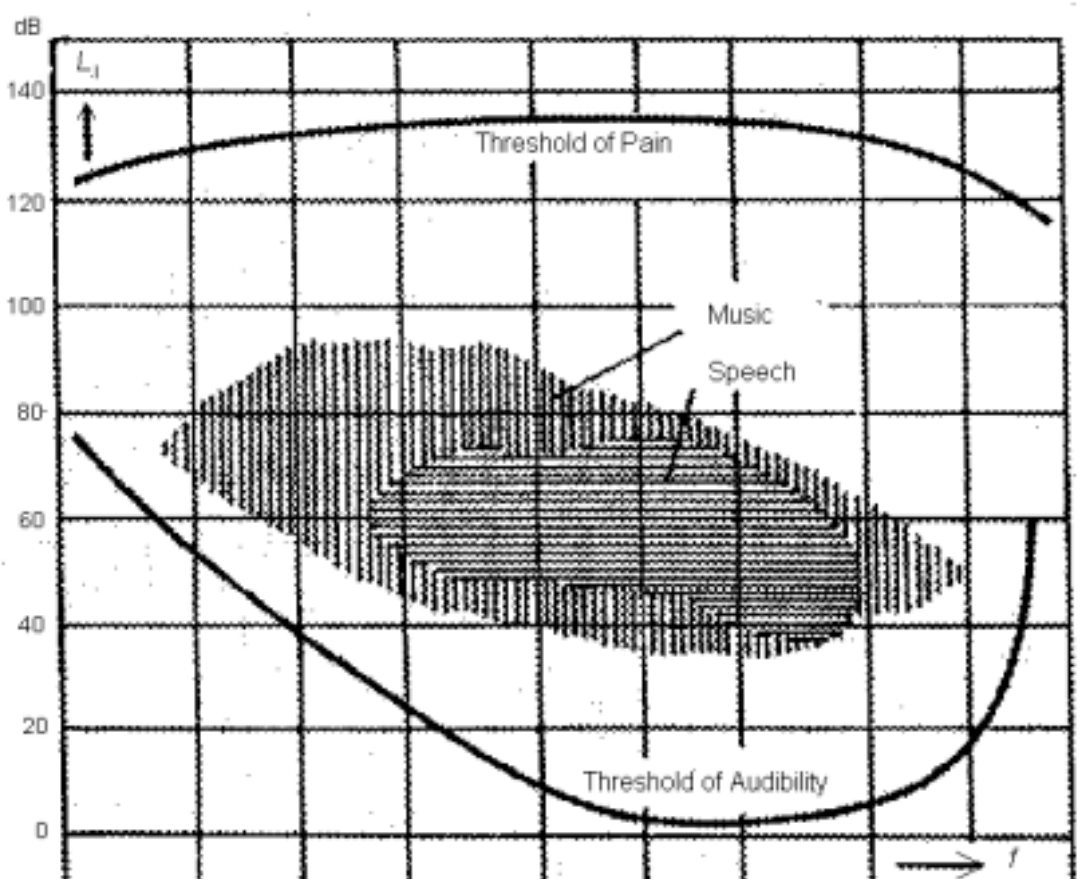


Figure 1.2.2. The frequency characteristics of human ears together with the typical frequency range of voice and music ($L_1 = 20\lg\frac{p}{20*10^{-6}\,p_a}$, p: effective voice presssure)

system. This effect is measured by the quantity $k = \frac{\sqrt{\sum_{i=2}^{\infty} c_i^2}}{\sqrt{c_1^2}}$ where, c_1 is the amplitude of the baseband sinusoidal with frequency f_0, while c_i is the amplitude belonging to the frequency if_0, *i=2,3....* Most of the analog broadcasting system requests *k*=1%, while in analog telephony the typical value of *k* is 5-10%.

Statistical properties of speech signal

Several studies have been dealt with revealing the underlying statistics of voice as a stochastic process, in order to develop optimal voice coding systems. The first issue to be tackled in the course of such an investigation is that the voice signal is highly nonstationary. This entails that short term statistics defined over the range of 32 ms fundamentally differs from long term statistics defined over the interval 6400

msec. Several authors tried to capture the long term probability density function of speech by using a Laplacian p.d.f. [1.2.1, 1.2.9], given as $p(x)=\frac{1}{\sqrt{2}\sigma}e^{-\sqrt{2}|x|/\sigma}$. The model probability density functions (p.d.f.-s) are indicated by figure 1.2.3.

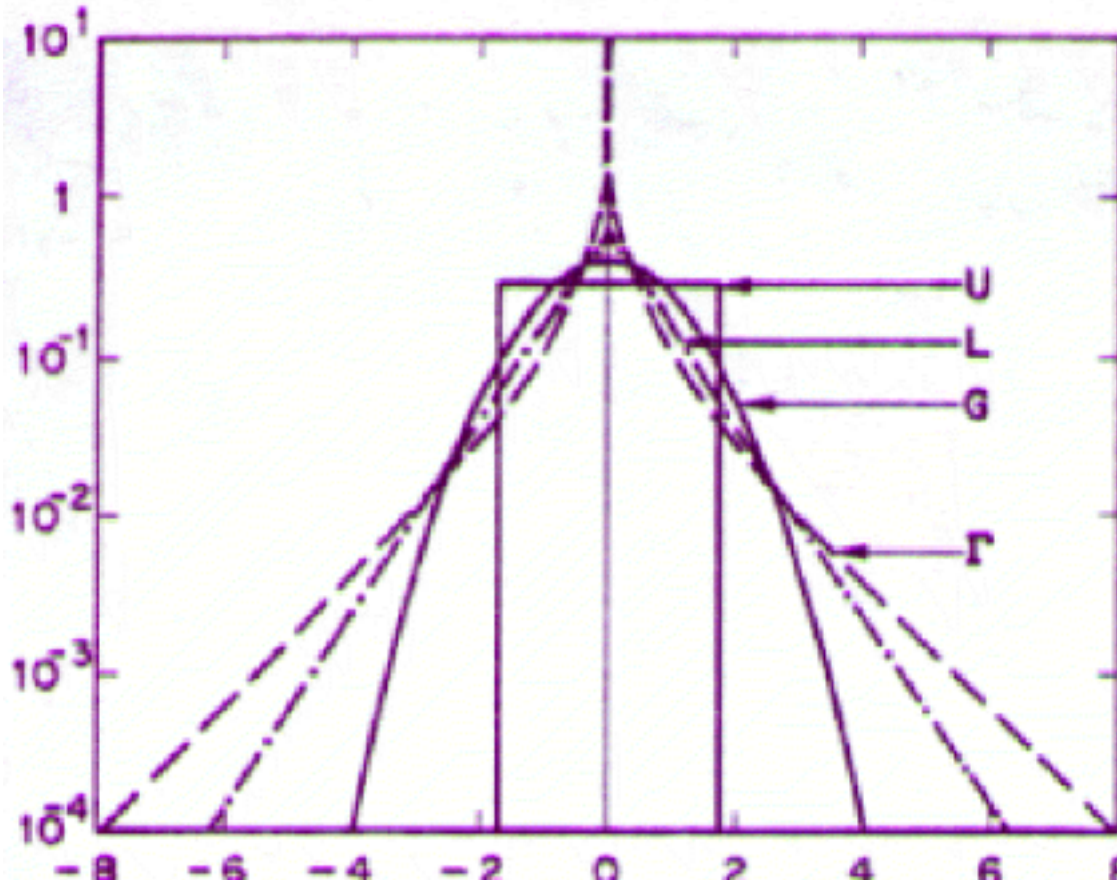


Figure 1.2.3. Four model p.d.f.-s for speech

The time dependence of short term speech variance [1.2.2] is shown by figure 1.2.4.

The signal to noise ratio of digital voice is defined as $SNR=10\lg\frac{\sigma_x^2}{\sigma_z^2}$, where $x_k, k=1,2,...$ are the samples of the original signals, while $z_k, k=1,2,...$ represent the quantization error.

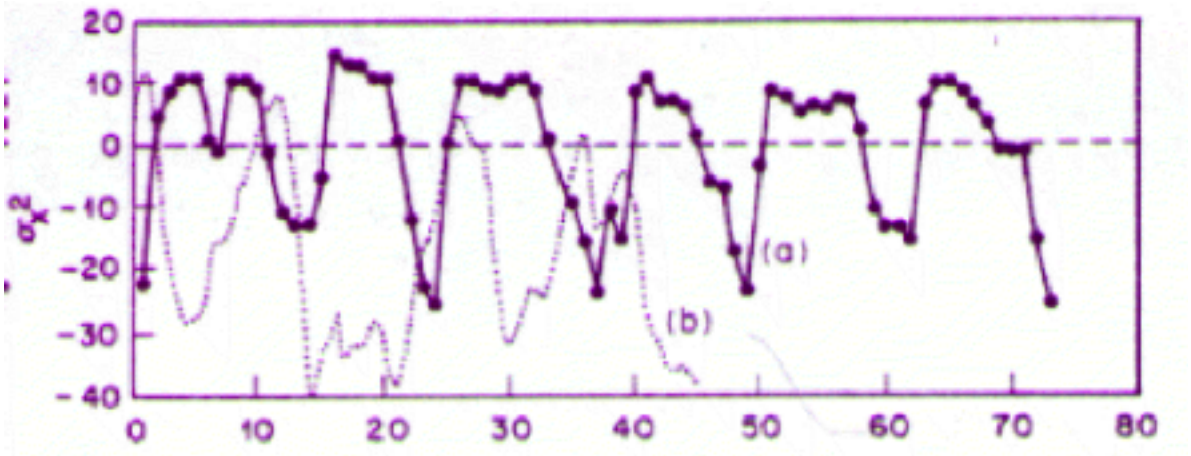


Figure 1.2.4. The logarithm of short term speech variance with respect to time

1.2.1.1. Digital telephony

Standard digital telephony (PCM systems) has defined a (0, 4KHz) bandwidth for speech and the PCM coder uses an 8KHz sampling rate for digitization followed by an 8-bit logarithmic quantizer. Thus, from digital point of view the necessary bandwidth is 64Kbps. One must note though, that this rate also depends on the quality (and therefore the algorithmic complexity) of the coder. For example, the so-called vocoder can provide higher quality voice signals than PCM coders thus, it needs less bandwidth but at the same time yields highly complex CODEC architectures. In PCM systems SNR increases very close to 6dB per additional bit [1.2.2], more precisely: $SNR = 6n - 10\lg a$, where *a* is an empirical constant in the range of 1<*a*<10, while *n* is length of binary codewords used in the quantizer.

For the CCITT 30 channel *A*-law system designed for multiplexing voice, each TDM frame is made up of a set of 32 channel time slots (0 to 31). Each channel time slots comprises 8 digit time slots a duration of 125microsec/32=3.9 microsec. Each digit time slot has a duration of 0.488 microsec.

For Voice over IP (VoIP) transmission, present technology uses CODECs labelled as (G.711, G.723, G.726, G.729, G.728). Depending on the applied compression technique, the associated data rate can vary between 6.4-64 Kbps. Smaller data rates need bigger digitizing delay and more powerful DSP architectures, though.

For mobile telephony, UNIX based implementation of GSM 06.10 full rate transcoding are available throughout Europe resulting in a 13 kbps bandwidth using long term prediction combined with residual pulse excitation.

1.2.1.2. Characterization of music

One can characterize music based on the same principles as was done with respect to speech. The related parameters are summarized in the next table:

Characterization of music	
Bandwidth	15 KHz (in FM radio), 20kHz (in CD)
SNR	40dB (in FM radio) , 96 dB (in CD)
Nonlinear distortion	1% (in FM radio), 0.005% (in CD)
Sampling rate	44,1 KHz (CD, MPEG)
Quantization	16 bit
Digital bandwidth	705.6 Kbps for monoaural, 1.411Mbps for stereo

1.2.2. Characterization of video information

Similarly to the description of audio information, in this section we summarize the properties of video information.

Perceptual considerations

Before getting down to the description of video signals, we have to briefly summarize the properties of human vision.

The eye can perceive light in the wavelength domain of 400nm to 700nm and its sensitivity is depicted by Figure 1.2.5.

Regarding colours, experiments has demonstrated that that most colours perceived by the eyes can be composed by Red ($\lambda_R = 700nm$), Green ($\lambda_G = 564.1nm$) and Blue ($\lambda_B = 435.8nm$) which is often referred to (RGB) representation. In this way every colour is uniquely defined by a three dimensional vector $\mathbf{x} = (x_R, x_G, x_B)$, where the components indicate what is the intensity of Red, Green and Blue in the corresponding colour. To simplify this description, the three dimensional colour space is mapped into a two dimensional representation called chromacity diagram. The contour in this chromacity diagram (apart from the bordering line below) represents the spectral colours, whereas the innermost region contains "muted" colours (spectral colour + white). The diagram is bordered by the so-called "purple line" containing the

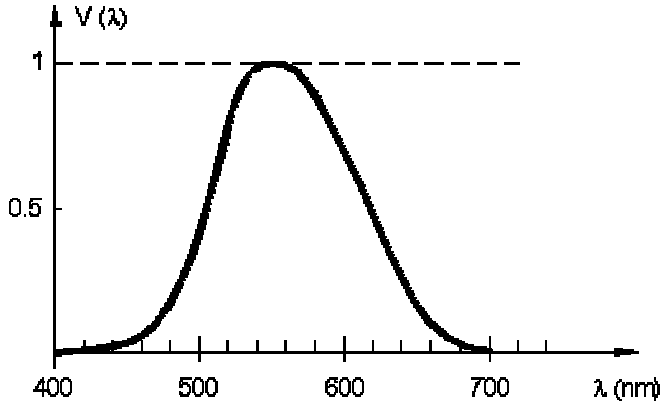


Figure 1.2.5. The visibility diagram

combination of red and blue colours. The coordinates of the white point *E* are (x=0.33 y=0.33).

The television signal

To introduce the television signal, first we have to take into account that the human eye can optimally focus on a rectangular image (having width/height ratio as 4/3) from the angle of 20 degree. Since the resolution of eye is 2', this yields that images must contain 800x600=480,000 pixels. There is a gray level associated with each pixel and the television signal contains the sequence of these gray levels, reading them out from left to the right, row by row. Theoretically this reading process is done at 25Hz rate, since human vision perceives the sequence of images as moving pictures at this frequency. However, in contrast to 25Hz rate of moving pictures, television adopts a 50Hz rate since the electric ray first projects the even rows and then the odd rows to the screen. Regarding the amplitude considerations, one can roughly estimate that the black and white television signal contains 70% video-information and 30% synchron information. In the case of colour TV the luminescence signal is defined as Y=0.3R+0.59G+0.11B, while the chrominance signals are *R-Y* and *B-Y*, respectively.

Historically, three major standards has evolved to code coloured video signals, NTSC, PAL and SECAM. NTSC transmits the chrominance signals by QAM thus it is very sensitive to phase distortion. SECAM takes advantage of the highly correlated video information and only transmits one chrominance signal at each row (the other

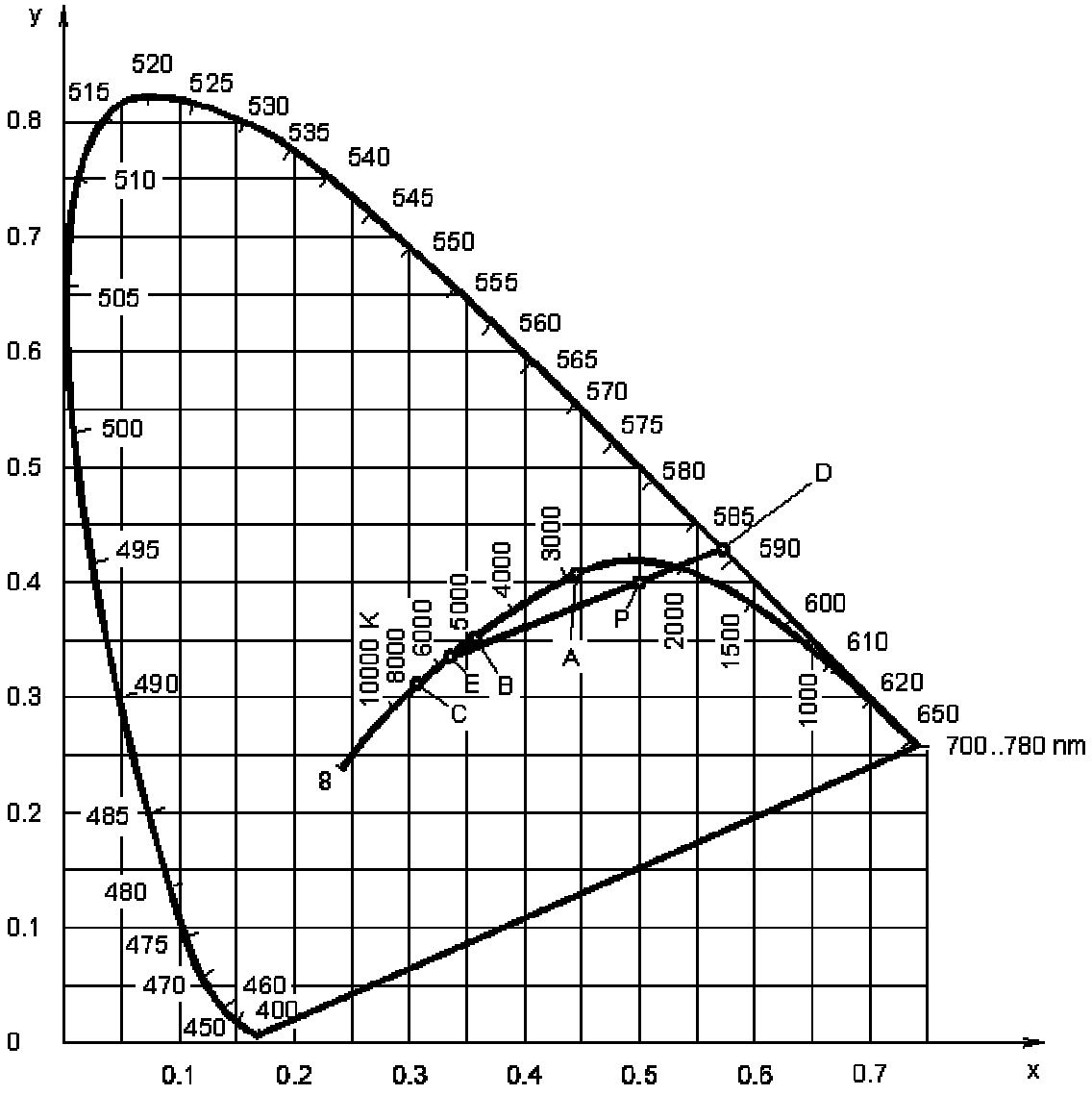


Figure 1.2.6. The chromacity diagram

one belongs to the previous row and is recalled from the memory). In this way, QAM is not needed (it is replaced with FM) which increased the quality of the system. PAL alternates the phase of the Red signal which improves the spectrum of the QAM signal in such a way that it less sensitive to phase distortions.

In the recent years, there has been a considerable interest for High Definition TeleVision (HDTV), which yields sharper images but roughly doubling the scan lines. In this case the size of the picture follows the ratio of 16/9 instead of 4/3 to match the format better to movies. It uses adaptive coding by using high resolution for still parts of the image while a lesser resolution for the moving part of the image. To further

utilize the bandwidth a so-called motion vector is transmitted instead of the motion itself (a more thorough description of HDTV can be found in [1.2.8]).

Digital bandwidth for transmitting video information

As was mentioned before, each image contains 800x600=48,000 pixels. There is a gray level associated with each pixel which can takes its value from 100 different quantized gray levels. As a result, one black and white image amounts to 48,000 ld 100 = 3.19 Mbit of data. Regarding moving pictures, the human eye can average standing images shown at the rate of 25 Hz into moving pictures. Therefore, the necessary bandwidth for BW moving pictures is $3.19 * 10^6 25 \approx 80 Mbps$

When coloured images are to be transmitted, we only have to deal with the two so- called chrominance signals as the luminance is already included in the BW image. Furthermore, we can rely on the fact that human eyes do not need such a fine resolution in the case of colour as in the case of BW images (the colour resolution is approximately one fifth of the BW resolution). This leaves us to transmit 166Kbit for colour information in the case of still images. Therefore, the bandwidth needed for transmitting coloured images are given as 25(3.19Mbit+0.166Mbit)=84Mbps. One can see, that there is not so much "overhead" in adding colour information to a BW image (the bandwidth is just a "little" higher).

Besides television, one must consider how to transmit pictures appearing on the screen of the computer. There are a couple of standards (640x480 VGA, 800x600 SVGA, 1024x768 XGA). An XGA display with 24 bits per pixel and 25 frames per sec needs 472 Mbps bandwidth.

1.2.2.1. Statistical properties of video signals

When describing video as random processes one is encountered by long range dependence meaning that the correlation pattern is considerably high even in the case large time lags between the corresponding values of the video signal. One of the reasons for this behaviour is that there are repeated pixel patterns, which tend to appear in each frame. Therefore, video signals have a quasi-periodicity at the frame frequency. Some typical correlation's is shown by figure 1.2.7.

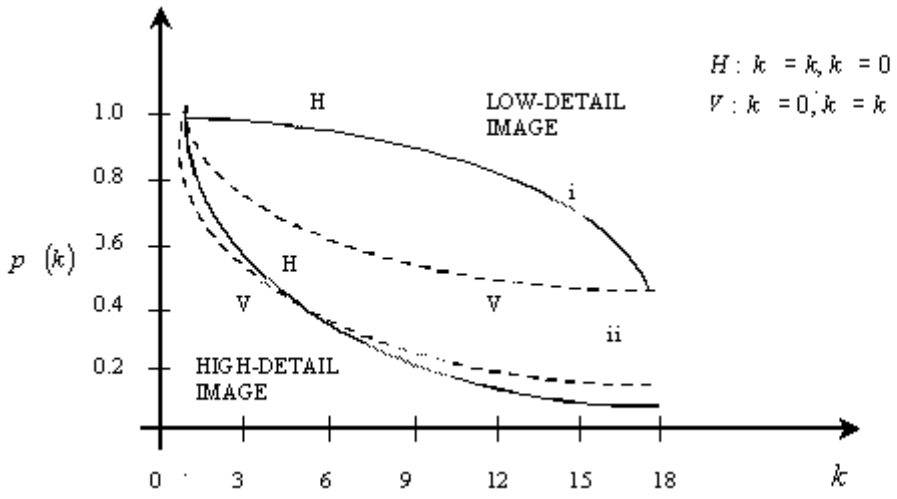


Figure 1.2.7. Temporal correlation video signal

One can see that the correlation of low detail image is very slowly decreasing in time indeed. For further details on the statistics of video signal see [1.2.4, 1.2.5].

1.2.3. Source modeling for data networks

In modern communication technologies every information process is ultimately transmitted via data networks. Therefore, it is important to characterize sources as either constant bit rate flows for circuit switched networks or as random packet streams for packet switched networks.

Source modeling for circuit switched networks

In the case of circuit switched networks resources are allocated permanently for a given a call. Therefore most of the services are offered a constant bandwidth. In this case information processes must be characterized by their bandwidth need and duration of time.

Figure 1.2.7 summarizes typical source processes in terms of required bandwidth and typical duration. For further details see [1.2.6].

Source modeling for packet switched networks

If information is transmitted via packet switched networks then modeling has to go one step further including traffic characterization of the source. This is necessary as information is transmitted on a packet by packet basis in contrast to

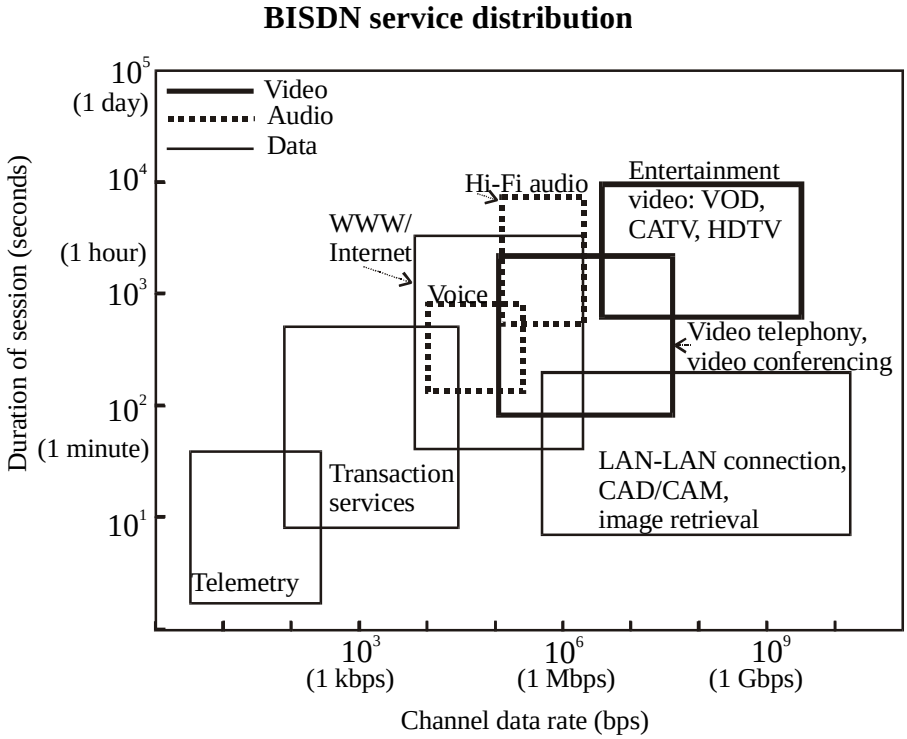


Figure 1.2.7 Typical bandwidths and duration of services in B-ISDN

constant bit rates. The importance of the task is demonstrated by the fact that traffic modeling itself has grown into a scientific discipline of its own, uniting methods from the fields of stochastic modeling, the theory of chaos ....etc. In this section we only describe some elementary models to capture the statistical nature of video, voice and data as packet streams. For further treatment and in-depth analysis see Chapter 1.7.

There are models with various depth modeling the statistics of information streams at the packet level, based on Markov Modulated Poisson Processes, fractal analysis...etc. Hereby, we only develop two type of models:

1. On/Off models for zero buffer approximation (in this case the packet switched network is dimensioned in such a way, that the bufferlength are considerably negligible in order to support real-time services, thus the major QoS concern is the Cell Loss Probability). In the case of On/Off modeling the underlying notion is to regard the source as a random variable denoted by *X* which needs bandwidth subject to a given probability distribution $p_i = P(X = i)$ where $i = 0,...,h$. Let $m = \sum_{i=1}^{h} ip_i$ denote the expected value of *X*. Usually the "burstiest" voice model is used often referred to as "On/Off" model substituting

the original distribution by a Bernoulli distribution given as $P(\tilde{X}=0)=1-\frac{m}{h}$ and $P(\tilde{X}=h)=\frac{m}{h}$, where $\tilde{X}$ is the bursty (or On/Off) equivalent of *X*.

2. Markov driven models which are essentially characterized by their states, state dwell times and state transition probabilities.

In the case of voice the typical parameters are m= 10Kbps, while h=64Kbps. Therefore, the aggregation of voice sources yield binomial distribution. It is also worth mentioning that the statistical bandwidth of On/Off voice sources, developed by Kelly to model the load traffic sources present to a packet switched network [1.2.3], is given as $\mu(s)=\frac{1}{s}\log\left(1-\frac{m}{h}+\frac{m}{h}e^{sh}\right)$ here *s* is a parameter subject to optimizing the Chernoff bound. The concept is based on using the logarithmic moment generating function to upperbound the tail of the aggregated traffic.

The Markov driven model for speech is given as shown in figure 1.2.8.

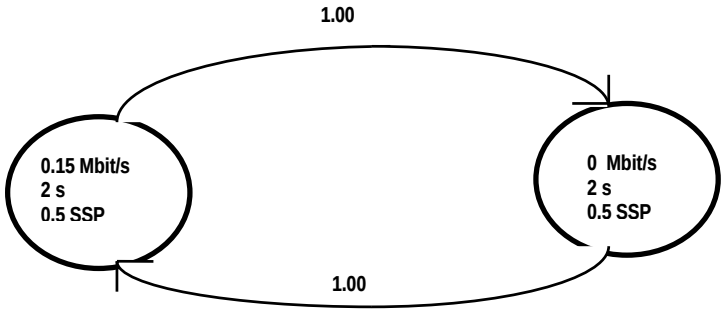


Figure 1.2.8. A Markov driven model of speech

Traffic modeling of video sources can also be done in an On/Off fashion, however, this does not reproduce the inherent correlation of video information. The corresponding On/Off model is given as follows:

m=2.2121 Mbit/s h= 11.02 Mbit/s

Another approach to model video-processes is to use Markov modulated Poisson Processes. This model, however, falls short of capturing long range dependence as well. Therefore video-processes, especially aggregated video poses a challenge to traffic modeling and management in packet switched networks.

The datafast model developed for *ftp* has the following characteristics m=0.4356 Mbit/s, h=11.01 Mbit/s

The Markov driven model [1.2.9] is given as follows:

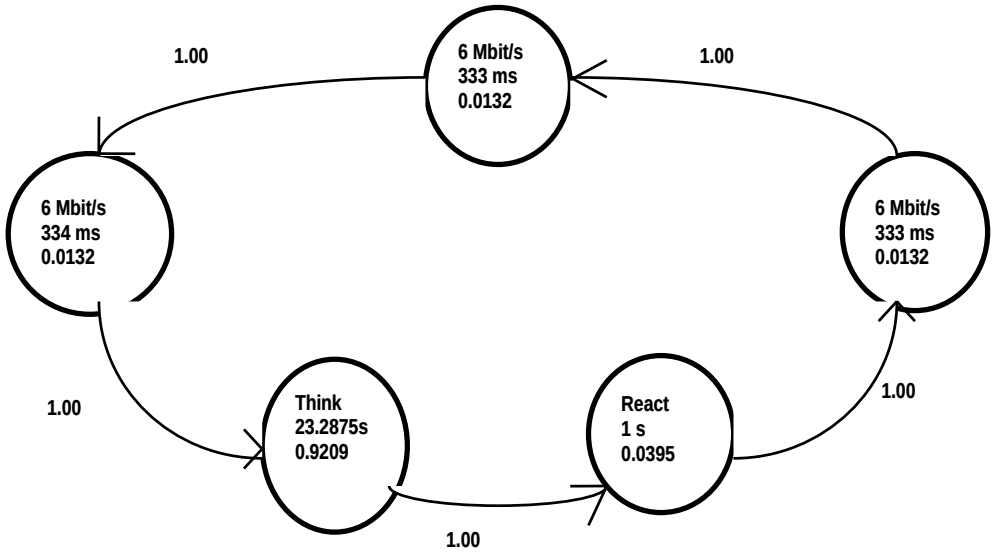


Figure 1.2.9. A Markov driven model for ftp

Summary

In this chapter the properties of real life information processes were investigated. Based on a layered approach, the attributes of analog voice, music and video were summarized followed by the characterization of voice, video and data sources as digital streams. Finally, packet level modeling was developed to capture the underlying statistical behaviour of sources.

References:

[1.2.1] Flanagan, J.L.: *Speech analysis, synthesis and perception,* Springer-Verlag, 1972

[1.2.2] Jayant, N.S. , Noll, P.: *Digital coding of waveforms*, Prentice Hall, 1984

[1.2.3] Kelly, F.: "Notes on effective bandwidths", *www.statslab.cam.ac.uk./~frank/*

[1.2.4] Kretzmer, E. R.: "Statistics of Television signals", Bell System Tech. J., pp. 751-763.

[1.2.5] Kummerow: "Statistics for efficient linear and nonlinear picture coding", Int. Telemetering Conf, pp.149-161, October 1972.

[1.2.6] Lee, B., Kang, M., Lee, J.: *Broadband telecommunications technology*, Artech House, 1996.

[1.2.7] Levendovszky, J., Elek, Zs., Végsõ, Cs.: "Validation of novel CAC algorithms", *ICAM- IEEE 1999*, pp. 195-211

[1.2.8] Nimoneya, et al.: "An HDTV Broadcasting System Utilizing a Bandwidth Compression Technique MUSE", *IEEE Trans. on Broadcasting*, pp 130-160, 1987

[1.2.9] Zelinski, R., Noll, P.: "Optimal quantization for PCM", Technical report , Heinrich Hertz Institute, Berlin, 1974 (in German)

1.3. Source coding and channel coding

István Vajda dr., author

József Dénes dr., reviewer

Communication systems transmit data from its source to the destination (see Figure.1.3.1.). Data is first processed by a source encoder which is designed to represent data in a more compact, compressed form. Encoders produce codewords at their output, where a codeword is a sequence from a given code alphabet. The *source codewords* appearing at the output are sequences of source codeword symbols. The output of the source encoder is fed into the channel encoder. This latter encoder transforms input sequences into *channel codewords*. In contrary to source encoders channels encoders expand the input in size by adding redundancy to their input. The modulator transforms the symbols of the channel codeword into signals (analog symbols) for transmission through the channel. In the channel signals are corrupted by various sources of noise, distortion and interference. Demodulator is the unit, which makes mapping between channel output signals and channel codeword symbols. The demodulated sequence of symbols corresponding to the symbols of the transmitted codeword is called the *received word*. The task of the channel decoder is to give the best estimate on the codeword and accordingly to calculate the estimated source codeword. Source decoder performs the decompression of this source codeword.

The aim is to transmit data as effectively as possible. This means, for example, that if the channel is given and the service quality is prescribed for data reconstruction at the destination, we would like to transmit data at highest speed possible.

Theoretical limits for source coding and channel coding are given by information theory. Among the results of the classical information theory we find also the first constructions for data compression. Most of practical methods in use for channel coding were provided algebraic coding theory. Naturally, a short chapter does allow only a nutshell presentation of main ideas, basic definitions from the field of these two large theories.

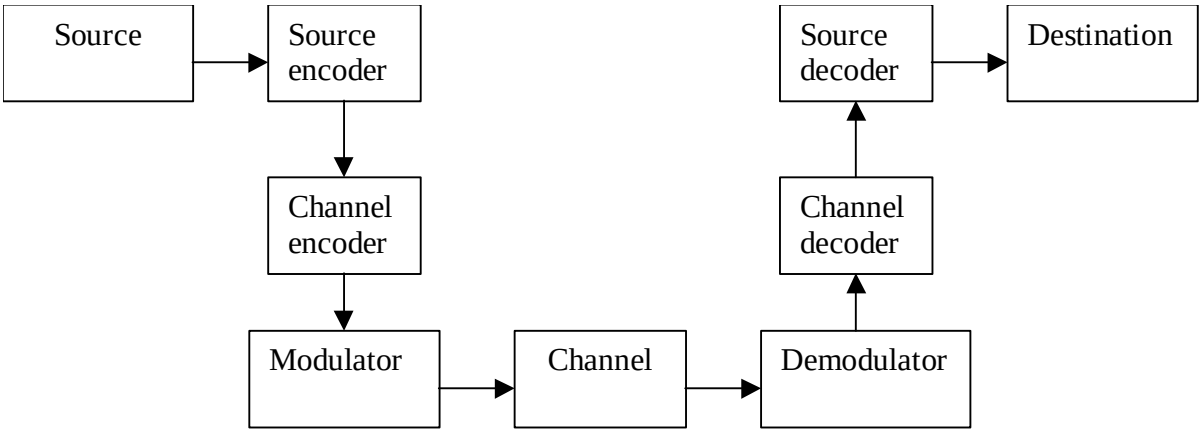


Figure 1.3.1 Block diagram of a digital communication channel

Data compression does not depend on the channel and the channel coding does not depend on the source distribution. It can be proved the the two-stage methodis as good as any other method of transmitting information over a noisy channel. Therefore algorithms can be developed separately. Below first we give a short introduction into source coding basics. Chapter 1.4. is devoted to data compression methods. Largest part of this chapter introduces channel codes.

1.3.1. Source coding: entropy and compression

The concept of information is too broad to be captured by a single definition. The entropy, which can be computed for any probability distribution, has many properties that agree with our intuitive notion about a measure of information.

Let X be a discrete random variable with alphabet X and probability distribution $p(x)=\Pr(X=x)$, $x \in X$. The *entropy* $H(X)$ of a discrete random variable X is defined by

$$H(X) = -\sum_{x} p(x) \log_2 p(x)$$

and it is measured in bits. For binary variable, when $P(x=1)=p$ and $P(x=0)=1-p$, this formula simplifies to

$$h(p) = p \log_2 p + (1-p) \log_2 (1-p)$$

In particular, the entropy of the binary outcome of the experiment of coin tossing is 1. The entropy is non-negative, because $0 \le p(x) \le 1$, it implies $\log_2 p(x) \ge 0$.

Assume that X has for outcomes: a,b,c,d, with probabilities 1/2, 1/4, 1/8, 1/8 , respectively. Person A selects an outcome according to this probability distribution.

The task for person B is to find out the outcome with minimum expected number of questions. He asks first: "Is it '*a*'?". If yes, then he stops, otherwise he asks: "Is it '*b*'?" . If yes, then he stops, otherwise he asks: "Is it '*c*'?" .

This way B resolved the uncertainty about the outcome with 3 binary questions. According to the given probabilities the average number of questions is: 0.5*1+0.25*2+0.25*3=1.75.

What is interesting here is that the entropy has the same value, i.e. $H(X)$=1.75! In general the minimum expected number of binary questions is between $H(X)$ and $H(X)$+1.

At this point we arrived at establishing the fundamental limit for the compression of information. Data compression can be achieved by assigning short descriptions (binary strings) to the most frequent outcomes of the data source and necessarily longer descriptions to less frequent outcomes. The famous Huffman coding procedure for finding minimum expected description length assignments is also explained below.

A *source code* C for random variable X is a mapping from range X to Y^*, the set of finite length strings of symbols from an s-ary alphabet, i.e. $Y^*=\{y_1,y_2,...,y_s\}$. Let $C(x)$ denote the codeword corresponding to x and let $l(x)$ denote the length of $C(x)$. The expected length $L(C)$ of a source code C for a random variable X with probability density $p(x)$ is given by

$$L(C) = \sum_x p(x)l(x) .$$

In the case of the above numerical example C={0,10,110,111}, where $C(a)$=0, $C(b)$=10, $C(c)$=110, $C(d)$=111, and $L(C)$=1.75.

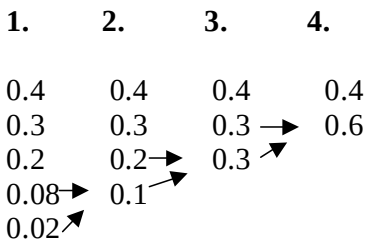
A code is called *prefix code* if no codeword is a prefix of any other codeword. The above code is prefix. If codewords are concatenated into a long string by person *A*, then person *B* can decode it (can cut the continuous long string into codewords uniquely) without reference to the future characters of the string since the end of a codeword is immediately recognizable. Therefore a prefix code is a *uniquely decodable code*.

For arbitrary uniquely decodable code C, the expected codeword length cannot be smaller than $H(X)/\log_2 s$. Fortunately the algorithm for finding optimal

uniquely decodable code is known: it is the Huffman code. Applying this code the expected codeword length between $H(X)$ and $H(X)+1$:

$$H(X)/\log_s s \le L(C) < H(X+1)/\log_s s + 1.$$

A numerical example is given below, for variable X with probability distribution: {0.4, 0.3, 0.2, 0.08, 0.02}.



We combine the two least likely symbols into one symbol, until we are finally left with only one symbol, and then we assign codewords to the symbols: we step backward starting at step 4.. First we assign codeword $C(x_1)$=0 to probability 0.4 and prefix 1 to all other probabilities of the distribution. This prefix is lengthened by bit 0 and 1, resulting in code $C(x_2)$=10 for probability 0.3 instep 3. The other fraction of 0.6 obtains prefix 11. In step 2. prefix 11 is lengthened with 0 and 1, and we obtain $C(x_3)$=110. In step 1. the prefix 111 is lengthened resulting in codewords $C(x_4)$=1110, $C(x_5)$=1111.

The advantage of self-synchronization capability of a prefix code is at the same time its disadvantage when they are transmitted through a channel, which errs. It may happen that losing synchronism at a point in a string results in wrong decoding of all future codewords. If instead we use code where the codewords have constant length, this problem can be eliminated. Unfortunately a uniquely decodable code with constant length also loses the capability for compression. Therefore we have to give up unique decoding, if we want constant codeword length. We accept distortion, introduced by such compression. Distortion is not acceptable when we compress, for instance, program codes, however it is acceptable for voice or image sources within the prescribed quality of service (see also chapter 1.4.)

1.3.2. Channel coding

Elementary concepts

Suppose that data is represented as binary information. A *binary code* of size *M* and block length *n* is a set of *M* binary words called - channel - *codeword*s. Usually, $M = 2^k$ for an integer *k*, which corresponds to the case when a block of k source code bits, called *message* is encoded into a codeword. The code is referred as an (*n*,*k*) binary code. For instance the (5,2) binary code

C={ 00000, 10101, 01111, 11010} (1)

consists of four codewords, where pairs of source code bits are encoded into channel codewords. In general, we can define codes over an arbitrary finite alphabet with *q* symbols {0,1,2,...,*q*-1}. Case *q*=2 is called binary.

There are two basic classes of channel codes: block codes and convolutional codes. The above-introduced code is a *block code*: the stream coming out of the source (source encoder) is divided into message blocks of size *k*, and messages are transformed into codeword blocks. The output block depends only on the input block. The *rate R* of a block code is defined by $R = k/n,\ 0 \le R \le 1.$ Higher rate corresponds to less redundancy. *Convolutional codes* are more complex. The binary input stream to the encoder is cut into short segments of length k_0, called input frame. At one step an input frame is entered into the encoder who produces an output frame of length n_0. The encoder has a state (memory), and its state and the actual input frame determine the content of the recent output frame. The rate *R* of a convolutional code is defined by $R = k_0/n_0,\ 0 \le R \le 1.$

Block codes are compared by three basic parameters: the block length *n*, the message length *k* and the *minimum code distance* d_{min}. Accordingly, we refer them as $C(n,k,d_{min})$ over code alphabet of size *q*.

The *Hamming distance* $d(x,y)$ between two codewords is the number of positions in which they differ. The minimum code distance is the smallest Hamming distance of the pairs of codewords. Considering code (1), d_{min}=3.

Assume that codeword *c* was transmitted and word r was received:

$$(r_0, r_1, ..., r_{n-1}) = (c_0, c_1, ..., c_{n-1}) + (e_0, e_1, ..., e_{n-1})$$

where the difference between r and c is called the *error vector*. For example if codeword 01111 was transmitted and word 01011 was received then a single error occurred, and e=(00100). The error occurred in position 2 and has value 1.

The two basic uses of channel codes are *error detection* and *error correction*. Error detection means that the receiver would like to detect the received word is in error, i.e. differs from the transmitted codeword. In case of a detected error the receiver asks for repeated transmission of the codeword, until it arrives error-free. The aim of error correction is to eliminate errors without relying on retransmission.

It is not hard to see that if a code has minimum distance d_{min}, then it has the capability to detect $t_{\det} = d_{\min} - 1$ errors: if a transmitted codeword is changed in less number of positions than its distance to the closest codeword, the received word will differ from any codeword, therefore it can be detected. In the case of code (1) we can detect any two errors.

Similarly if a code has minimum distance d_{min}, then it has the capability to correct $t_{corr} = \text{int}[(d_{\min} - 1)/2]$ errors, where function *int*(x) gives the largest integer smaller than x. Until the received word is closer to the transmitted codeword than to any other codewords, searching for the codeword which is at minimum distance to the received word, the transmitted codeword can be found successfully.

Considering a non-binary code it can happen that we know the position of errors but we do not know the value of error. Such an error is called *erasure*. If code has minimum distance d_{min}, then it has the capability to correct $t_{eras} = d_{\min} - 1$ erasures: if a transmitted codeword is erased in less number of positions than its distance to the closest codeword, the portion of codeword made up of non-erased positions will best fit to the transmitted codeword. In the case of code (1) we can correct two erasures.

The simplest code is the *simple parity-check code*, $C(n, n\text{-}1, 2)$, q=2. A single parity bit is added to the message. This code is able to detect a single error. For instance code C={ 0000, 0011, 0101, 0111, 1001, 1010, 1100, 1110} has parameters $C(4,3,2)$.

The simplest error correction code is the *simple repetition code*, $C(n,1,n)$, q=2. This code consists of only two codewords the all zero and all one codeword. The

message is one bit: if its is 0 the all zero codeword is transmitted, otherwise the all one codeword is transmitted. For instance code *C*={ 00000, 11111} has parameters *C*(5,1,5), and can correct 2 errors or detect 4 errors. The rate is low: *R*=1/5.

The *2-dimensional parity check code* is also elementary, but has improved efficiency. The message is arranged into a matrix. Its rows and columns are lengthened by a corresponding parity bit: consider, for instance, 3x3 bit size message matrix

1	1	0	*0*
0	1	1	*0*
1	1	1	*1*
0	*1*	*0*	*1*

and add a 4th row and a 4th column of parity bits. The bit at the lower right corner is called parity of parities, which can equally be calculated as the parity of the row-parities or the parity of column parities. This code has distance 4, consequently it is able to correct one error. The correction algorithm is also very simple: the indices of the row and column with parity error point to the entry in error.

Nonbinary arithmetic

The modern algebraic codes extensively use tools from algebra. One of the most important tool is connected to the finite fields (Galois-field). A field is a set F with multiplication and addition operations that satisfy the familiar rules - associativity and commutativity, the distributive law, existence of an additive identity 0 and a multiplicative identity 1, additive inverses, and multiplicative inverses for everything except 0.

The following fields are basic in many areas of mathematics: (1) the field Q consisting of all rational numbers; (2) the field R of real numbers; (3) the field C of complex numbers; (4) the finite field GF(q). GF(q) is an algebraic structure with q elements. The number of elements can only be a prime number or a power of a prime number: $q = p^m$.

The simpler case is when q=p, i.e. m=1. For instance, consider GF(5). There are 5 elements in this field: 0,1,2,3,4. The addition and multiplication is done modulo 5, for instance, 3+4=7=5+2=2 mod 5. The inverse of 4 is 4, i.e. $4^{-1}=4$ because

4*4=16=3*5+1 mod 5. The simplest case is GF(2), with the usual modulo 2 arithmetic. Field *GF*(*p*) can be extended to the general case of m>1:

A polynomial over *GF*(*p*) is a mathematical expression

$$f(x) = f_{n-1}x^{n-1} + f_{n-2}x^{n-2} + ... + f_1 x + f_0$$

where the symbol x is an indeterminate, the coefficients f_{n-1},..., f_0 are elements of *GF*(*p*) and the indices and exponents are integers. A polynomial that is divisible only with $\alpha p(x)$ or α, $\alpha \in GF(p)$ is called irreducible polynomial. A monic polynomial is a polynomial with leading coefficient f_{n-1} equal to 1. A monic irreducible polynomial of degree at least 1 is called a prime polynomial. Consider the ring of polynomials modulo *p*(*x*), i.e. where the two operations are the usual multiplication and addition operations of polynomials modulo *p*(*x*). For an arbitrary polynomial *p*(*x*) over *GF*(*p*), such a ring satisfies all rules mentioned above for fields except the existence of multiplicative inverses. However it can be shown that:

The ring of polynomials modulo a monic polynomial *p*(*x*) is a field if and only if *p*(*x*) is a prime polynomial.

Consider, for instance, Galois-field $GF(2^3)$ (=*GF*(8)). The elements of the field, denoted by 0,1,2,3,4,5,6,7, are represented by binary polynomials of degree less than 3, i.e. by zero, one and two degree polynomials:

GF(8)={0, 1, x, x+1, x^2, x^2+1, x^2+x, x^2+x+1} ,

where we can use natural correspondence {0 <--> 0, 1 <--> 1, 2 <--> *x*, 3 <--> *x*+1, 4 <--> x^2, 5 <--> x^2+1, 6 <--> x^2+x, 7 <--> x^2+*x*+1} according to relationship between the coefficients of polynomial and the binary representation of the number-labeled elements. The operations are done modulo x^3+x+1, where the latter polynomial is selected as a binary irreducible polynomial with degree 3. The addition is easy: for instance, 3+5=(*x*+1)+(x^2+1)=x^2+*x*=6. In fact, the multiplication is done modulo x^3+*x*+1: for instance 3*6=(*x*+1)*(x^2+*x*)= x^3+2x^2+*x*= x^3+*x*=(*x*+1)+*x*=1 mod x^3+*x*+1, consequently

3^{-1}=6.

The *q*-ary alphabet for channel encoders is taken *GF*(*q*) .

Linear block codes

If any linear combination of codewords from a code $C(n,k,d_{min})$, $GF(q)$ results in codeword from the same code, we call code C linear. Shortly a linear block code is *linear subspace* of linear space of all q-ary vectors with length n. There are q^k codewords, therefore this code as a linear subspace has dimension k.

Matrices can describe linear codes: the *generator matrix* and the parity check matrix. The generator matrix G has dimension kxn, where the rows are linearly independent codewords (a set of vectors is linearly independent if nonzero linear combination cannot result in all zero word). Its rows correspond to the base vectors for spanning the code, i.e. producing the codewords:

$$(c_0, c_1, ..., c_{n-1}) = (m_0, m_1, ..., m_{k-1}) \cdot G$$

where the components of message vector $m = (m_0, m_1, ..., m_{k-1})$ linearly combine the rows of G, which results in codeword c.

The *parity check matrix H* has dimension (n-k)xn. It has the following property: for arbitrary codeword c from code C: $Hc^T=0$, where c^T denotes the transpose of row vector c, which means that codewords and only the codewords are orthogonal to the rows of matrix H. Matrix H plays role in decoding of linear codes as it will be explained below.

A code where codewords have the structure: $c = (m_0, m_1, ..., m_{k-1}, p_0, p_1, ..., p_{n-k-1})$, i.e. message appears clearly as portions of the codeword, is called *systematic*. The other portion of the systematic codeword is the parity portion

Systematic codes are practical: after when the decoder has decided on the transmitted codeword, the calculation of the corresponding message is trivial, simply cutting the prefix of the decoded codeword.

Considering code (1) it is easy to check that

$$G = \begin{pmatrix} 10101 \\ 01111 \end{pmatrix} \text{ and } H = \begin{pmatrix} 11100 \\ 01010 \\ 11001 \end{pmatrix} \qquad (2)$$

furthermore the generation is systematic.

It is not hard to see that for linear codes the minimum code distance is equal to the minimum Hamming weight of nonzero codewords. This follows from three simple facts: by linearity, the difference of any two codewords is a codeword; the Hamming distance is the Hamming weight of the difference; any codeword can be described as a difference: by simply subtracting the all zero codeword from it. This important property simplifies the calculation of minimum distance.

It is easy to give a proper lower bound on the minimum code distance according to a simple algebraic property of the parity check matrix: the minimum number of linearly dependent columns gives lower bound on d_{min}. Considering matrix H under (2) above, we can see that arbitrary two columns are linearly independent, and we need at least three of them, which combine to zero: for instance the sum of the columns 0,2 and 4 give zero. Therefore the minimum code distance is at least 3 (which we know exactly 3).

According to this theorem below we give the parity check matrix of the well know Hamming code $C(7,4,3)$ GF(2):

$$H = \begin{pmatrix} 0111100 \\ 1011010 \\ 1101001 \end{pmatrix} \qquad (3)$$

Cyclic linear block codes

A linear code $C(n,k,d_{min})$, $GF(q)$ which has the property, that arbitrary rotation (cyclic permutation) of a codeword is a codeword from the same code, is called *cyclic*. Cyclic linear codes are best described by polynomials. We make the following unique correspondence:

$$c = (c_0, c_1, ..., c_{n-1}) \; <---> \; c(x) = c_0 + c_1 x + ... + c_{n-1} x^{n-1}$$

Cyclic linear codes can be described by polynomials: the *generator polynomial,* $g(x)$ and the *parity check polynomial*, $h(x)$. Polynomial $g(x)$ has degree n-k, polynomial $h(x)$ has degree k. Using $g(x)$ we can generate codewords in polynomial form:

$$c(x)=m(x)g(x) \qquad (4)$$

where $m(x) = m_0 + m_1 x + ... + m_{n-1} x^{n-1}$ is the message in polynomial form. Between polynomials $g(x)$ and $h(x)$ we there is a simple relationship: $g(x)h(x)=x^n-1$. The well known *cyclic linear Hamming code* with parameters $C(7,4,3)$ $GF(2)$ has generator matrix $g(x)=x^3+x+1$. For instance, message $m=(1011)$ <---> $m(x)=1+x^3+x^4$ is encoded into codeword $c(x)=1+x+x^5+x^6+x^7$ <---> $c=(1,1,0,0,0,1,1,1)$. Generation (4) is non-systematic. The systematic generation can be obtained the following way: calculate the polynomial remainder

$$r(x) = - x^{n-k} m(x) \mod g(x), \quad (5)$$

which gives the wanted systematic codeword as

$$c(x)= x^{n-k} m(x) + r(x). \quad (6)$$

Generation (6) is also the mathematical form of CRC (*Cyclic Redundancy Check*) generation: polynomial $g(x)$ is called CRC generator polynomial.

Reed Solomon codes

Reed-Solomon codes are efficient nonbinary cyclic linear block codes. The generator polynomial of Reed-Solomon code $C(n,k,d_{min})$, $GF(q)$, $n=q-1$ can be given by:

$$g(x) = (x-\alpha)(x-\alpha^2)...(x-\alpha^{n-k-1}) \quad (7)$$

where $\alpha \in GF(q)$ has order n, which means that n is the smallest positive integer for which $\alpha^n = 1$. Reed-Solomon codes are optimal, they are MDS (Maximum Distance Separable) codes:

$$d_{min}=n-k+1$$

which is the maximum distance possible, when parity length $n-k$ is given. For example, a Reed-Solomon code $C(12,8,5)$, $GF(13)$ can be generated by:

$$g(x) = (x-2)(x-4)(x-8)(x-3),$$

which code can correct two errors. A practical parameter setting is $C(255,255-2t,2t+1)$, $GF(256)$, where t is the wanted correction capability, and the code uses byte alphabet.

Decoding concepts

The decoder tries to minimize the probability of erroneous decoding. The optimal approach is the *maximum likelihood (ML) decoding*. Assume we know a probabilistic model of the channel, from which we can calculate the conditional probabilities P(*r*|*c*) for arbitrary codeword *c* and received word *r*. Assuming that *r'* has been received, we search for a codeword *c'*, which maximizes the conditional probability P(*r'*|*c'*). Codeword *c'* is the ML-decoded codeword. Unfortunately for practical channel such a complete and accurate probabilistic description of the channel is not available. Furthermore even if we knew description of the channel, the search for the maximum is not feasible for large - practical size - codes.

The decoders for block codes usually follow an other approach: they try to find a codeword which is the nearest to the received word in Hamming distance, i.e. *minimum Hamming-distance decoding rule* is applied. Straightforward search for the nearest codeword is not a feasible approach for typical code sizes. *Syndrome decoding* provides the basic solution. The syndrome s for received word *r* is defined by

$$s=Hr^T.$$

If there were no errors, i.e. when *r*=*c*, we obtain *s*=0, as it follows from the basic property of matrix *H*. The correctable errors and the corresponding syndromes are enlisted - in principle - in a table, the syndrome decoding table. For the correctable error a vector with the smallest weight is selected from the set of possible vectors with the given syndrome. For our code (1) according to parity check matrix (2) the syndrome decoding table is the following:

s	e
000	00000
001	00001
010	00010
011	01100
100	00100
101	10000
110	01001
111	01000

Assume that the received word has syndrome 111. We look up the table and read error pattern 01000, i.e. we decide on the codeword which is obtained from the

received world by inverting its second bit. For usual code sizes working with syndrome tables is not feasible (the number of rows of this table is 2^{n-k} for a binary code). Therefore we use syndrome decoding algorithms instead, which means we calculate the error pattern again and again for each received word. The classical algorithm for this task is the Peterson-Gorenstein-Zierler-algorithm (PGZ), the more practical improvement is the Berlekamp-Massey algorithm. These algorithms set up and solve systems of linear equations where the unknowns are the position of the error and the value of the error, while syndromes are the known quantities.

Convolutional codes

The encoder of a convolutional code has a simple building elements, shiftregisters and mod 2 adders:

The encoder in Figure1.3.2. outputs two channel bits for an input data bit, i.e.

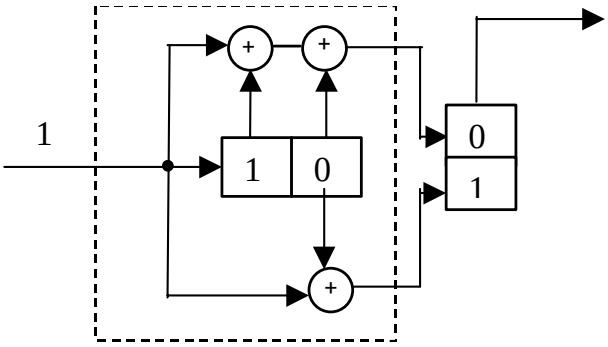


*Figure.1.3.2: Convolutional encoder [**x**²+**x**+1 , **x**²+1]*

has code rate R=1/2. The input bit steps into a shift register of 2 cells. Despite of the block codes the length of the input is not fixed, i.e. the length of the output sequence (codeword) is simply the double of the length of the input sequence (message).

The bits stored temporarily in the register are added modulo 2 (XOR) with the input bit. The generator is specified by its two generator polynomials corresponding to the positions of taps of the register.

The convolutional code is linear, which means that if codewords c_1 and c_2 correspond to messages m_1 and m_2, respectively, then the sum of codewords corresponds to the sum of messages. We have seen that for linear codes the minimum distance between codewords is equal to the minimum nonzero weight. If all zero input is fed into the convolutional encoder all zero output is obtained. In the case

of the generator in Figure1.3.2., the minimum weight codeword is obtained if the initial state of the register is 00 and the input is 1000... .The output becomes 11 10 11 00 00...., consequently the minimum distance (called also free distance) is 5.

For the convolutional codes there is feasible implementation of a maximum likelihood decoder, the Viterbi-algorithm. The code is represented by a special graph, called trellis:

The nodes correspond to the states of the shift register. Furthermore the paths and branches correspond to codewords and one step outputs of the encoder, respectively. These branches are labeled by the actual outputs of the encoder.

If the channel can be described by a discrete memoryless binary symmetric channel (DM-BSC) model, which means that channel bits are affected independently

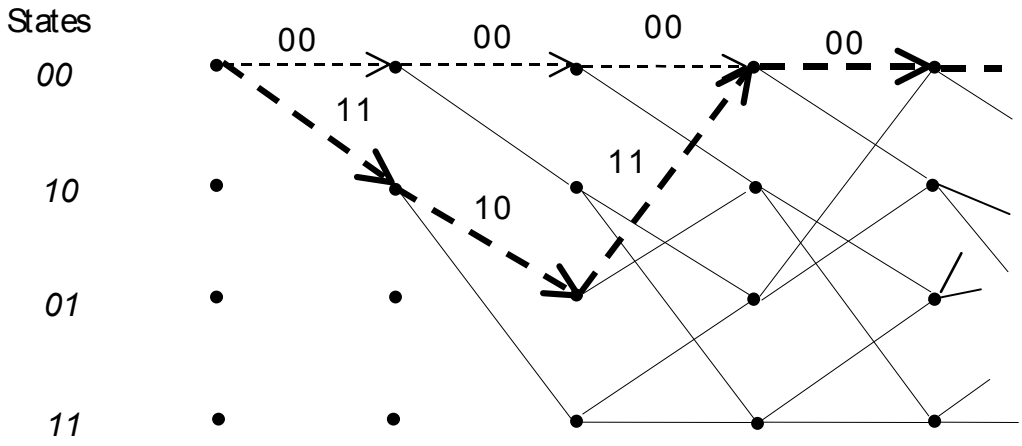


Figure 1.3.3: The trellis of a convolutional encoder

and the probability of 0-->1, 1-->0 errors are the same, then the maximum likelihood decoding is equivalent to minimum Hamming-distance decoding. Assume all zero codeword has been transmitted and 10 01 11 00 00 ... has been received.

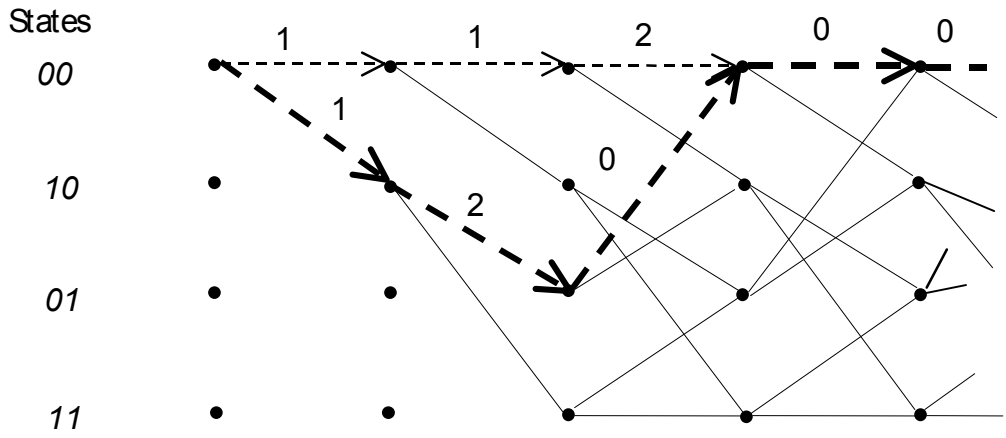


Figure 1.3.4: Trellis with weighted branches (detail)

All the branches are labeled by weights, which is the distance of the branch from the corresponding fragment of the received. In the case of DM-BSC channel the weights are calculated from Hamming distance. A detail is shown in Figure 1.3.4. The weight of a path is the sum of weights of its branches. Path with minimum weight corresponds to the codeword closest to the received word. The Viterbi algorithm can efficiently find minimum weight path in a weighted trellis.

The weighting can be generalized to general memoryless channel: the weight of a branch labeled by bitstring x is $log[p(y|x)]$, where y is the corresponding fragment of the received words.

Spectral bit rate and bit energy to noise ratio

Consider the encoder and modulator as one unit, which accepts messages and outputs signals. In other words a message selects a signal from the set of signaling waveforms. Let the S denotes the average power of the signal sent into the channel. If R_b [bit/sec] stands for the bit rate then the bit energy is $E_b=S/R_b$. Shannon gave the capacity C [bit/sec] for additive gaussian noise channel:

$$C = W \log_2\left(1+\frac{S}{N_0 W}\right) \qquad (8)$$

where W is the bandwidth of the channel which means that signaling waveform $s(t)$ of power $S=E_bR_b$ has spectrum which is zero outside of band $[-W/2,W/2]$, furthermore N_0 [watts/Hz] denotes the one-side spectral density of the white noise signal. Formula (8) says that any signaling waveform can transmit no more bits per second than C

through an additive white gaussian noise channel at an arbitrarily low error bit rate. Therefore this rate is challenging and provides a reference against which a digital communication channel can be judged.

From (8) useful bounds can be derived easily. It is interesting what happens if we allow large bandwidth (in principle arbitrarily large). It turns out that the smallest bit energy to noise ratio E_b/N_0 must remain above -1.6dB if we want to reach low - in principle arbitrarily low - bit error rate.

Experience shows that it is easy to design waveforms that provide small bit error rate if the ratio E_b/N_0 is about 12 dB (common digital modulation without encoding). The effectiveness of encoders can be measured in *coding gain* [dB], which is the decrease in ratio E_b/N_0 to attain the same bit error rate. This way different designs of encoder-modulator units can be compared, which is usually done in the needed bit energy to noise ratio needed to attain a given bit error rate (say BER=10^{-5}), where the compared units are placed between the same source and channel. The most promising encoders today use *turbo codes*. Using these codes ratio E_b/N_0 is about 1 dB (!) is achieved at practical BER values. These codes are improvements on classical convolutional codes and use complex iterative decoding.

If we cannot keep the same bit rate and bandwidth when comparing designs we can calculate the spectral bit rate $r = R_b/W$ [bit/sec / Hz]. The spectral bit rate r and E_b/N_0 are the two most important figures of merit in a digital communication system. From (8) it also follows that for any transmission systems, which wants to reach arbitrarily low bit error rate

$$\frac{E_b}{N_0} \geq \frac{2^r - 1}{r} \tag{9}$$

furthermore the bit energy to noise ratio by appropriate design - in principle - can approach $f(r)=(2^r-1)/r$. A graphical illustration can be seen in Figure.1.3.5.

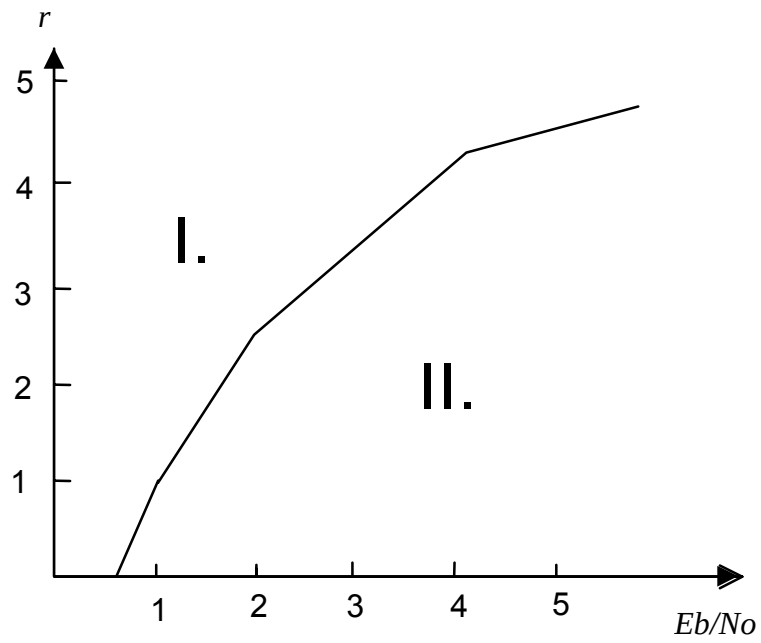


Figure 1.3.5. : Accessible spectral rates and bit energy to noise ratios

Region II. in Figure. 1.3.3. corresponds to accessible pairs [r, E_b/N_0] (scales are linear), i.e. for any point below the curve one can design a communication system that has as small a bit error rate as one desires. The border line of the region is given by function $f(r)$. For instance, if r tends to zero (W tends to infinity) the lowest possible bit energy to noise ratio is 0.69 (-1.6dB).

References:

[1.3.1] Györfi L, Györi S., Vajda I.: *Információ és kódelmélet*, Typotex Kiadó, 2000.

[1.3.2] Blahut, R.E.: *Theory and Practice of Error Control Codes*, Addison Wesley, 1986.

[1.3.3] van Lint,J.H.: Introduction to Coding Theory, Spinger Verlag, 1982.

[1.3.4] McElice, R.J. *The Theory of Information and Coding*, Addison Wesley, 1977.

[1.3.5] MacWilliams,F.J. and Sloane, N.J.A: *The theory of Error-Correcting Codes,*
Part I-II., Noth-Holland Publishing Co., 1977.

[1.3.6] Steele,R and Hanzo,L.: Mobile radio Communications, IEEE Press - John Wiley, 2000.

[1.3.7] Viterbi,A.J.: CDMA.Principles of Spread Spectrum Communication, Addison

Wesley, 1995

[1.3.8] Proakis,J.: Digital Communications, McGraw Hill, 1992.

[1.3.9] Simon,M. et.al.: Spread Spectrum Communications, Computer Science Press,

Vol I-III, 1985.

[1.3.10] Shannon,C.E.: *A mathematical theory of communication* ,Bell System Tech.

J., Vol. 27, 1948, pp. 379-423 (Part I), pp. 623-656 (part II).

[1.3.11] Berrou,C., Glavieux,A. and Thitimasjshima,P: *Near Shannon limit error-*

correcting coding and decoding:Turbo codes', Proceedings IEEE Int.Conf. on Communications, Geneva, May 1993, pp.1064-1070.

[1.3.12] Biglieri,E., Proakis,J and Shamai,Sh.: *Fading channels: information-theoretic and*

communication aspects, IEEE Trans. Information Theory, Vol. 44, No. 6, October 1998, pp. 2619-2692.

[1.3.13] Forney D. and Ungerboeck,G: *Modulation and coding for linear Gaussian channels*,

IEEE Trans. Information Theory, Vol. 44, No. 6,October 1998, pp. 2384-2415.

1.4. Data compression

János Levendovszky dr., author

György Dallos dr., reviewer

This chapter deals with modern data compression algorithms which aim at achieving considerable bandwidth reduction in the data to be transmitted. There are plenty of motivations to compress data regarding storage space and processing speed. In telecommunications, however, one of the primary motivations is to reduce bandwidth in order to transfer high-speed information via narrowband channels. This problem has come to the fore in the advent of widespread networking, where multimedia dataflows are downloaded from the web via narrowband access networks, which were originally designed to support voice communication. Therefore, to facilitate broadband information transmission and efficient bandwidth utilization in general, it is mandatory to develop novel data compression techniques.

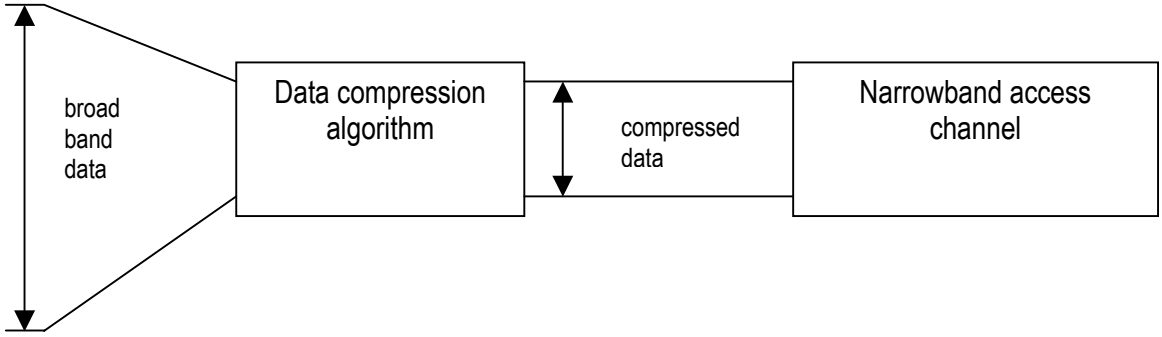


Figure 1.4.1 The aim of data compression

To grasp the importance of the task, one can take into account that data compression - which was traditionally labelled as part of signal processing and information theory - has now developed into an "independent" discipline pursued by many well-known scientist. Furthermore, beyond the scope of scientific journals, several monographs and books are dedicated to the subject [1.4.3, 1.4.6, 1.4.7] and innovative scientific ideas has already been matured into standards, such as JPEG, GnuZIP, MPEG ....etc.

Data compression can be approached from two different angles:

1. If data is given as a sampled analog signal then one can use transformation and adaptive predictive techniques to reduce the bandwidth. These methods are often referred to as "lossy methods" since in the course of transformation some "controlled" loss of information occurs. This implies that the original information cannot be fully restored from the compressed version. In spite of this side effect, these techniques are widely used as they can achieve high compression rates. For further details regarding the fundamental result see [1.4.6, 1.4.7]

2. If data is given as a bit stream then one can use traditional information theoretical methods to decrease redundancy and achieve the ultimate, shortest representation of signals. These methods are either rooted in adaptive source encoding schemes (e.g. Huffman coding) or in dictionary based methods like the Lempel-Ziv algorithms. Since the source information can fully be regained in this case, the corresponding methods are referred to as "lossless schemes". For a thorough survey in the literature, the readers are kindly referred to [1.4.1, 1.4.3, 1.4.5].

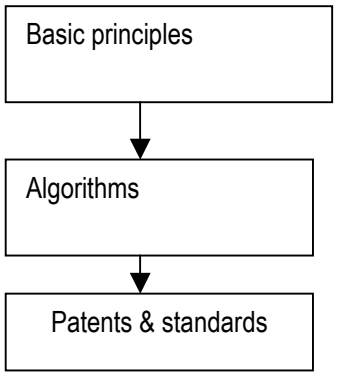
In the case of digital communications any type of information process is finally represented by a bit stream (let it be originally voice or a file generated by a word processor), thus the second methodology is universally applicable. Nonetheless, it may not fit to compress analog information (voice, images) because it cannot achieve a reasonably high compression rate. Therefore, lossless encoding schemes are typically applied to binary data (e.g. to files generated in binary formats or to digitized forms of analog information after lossy encoding has already been carried out).

Another factor to be taken into account when dealing with data compression is the algorithmic complexity often labelled as the "hidden cost". This cost is determined by two factors:

- what amount of information is needed to carry out the de-compression algorithm at the receiver side (e.g. a full dictionary has to be transferred in the case of dictionary based methods);
- what is the underlying algorithmic complexity of compressing and de-compressing algorithms in terms of number of operations ...etc.

The last factor can be critical when data "hits" the coder in full speed and compression must be done "on the fly". In this case it is mandatory to use computationally simple algorithms which can run on commercially available processing units. On the other hand, when data is available for leisurely perusal (e.g., in the case of still images) then more sophisticated compression algorithms are promoted to be used.

Our treatment of data compression methods touches upon principles, algorithms and standards, therefore the related material will be presented in the following order:



1.4.1. Basic principles

The underlying idea of compressing data comes down to exploiting the inherent statistical dependence being present in the source information. When pronouncing the beginning of a word or a sentence, the beginning often determines the end. Thus in speech or written documents it is easy to predict what will follow. In images there are many recurring patterns and periodicity which one could get rid of while compressing data. As a consequence, compression becomes possible because of highly correlated information flows. The efficiency of compression algorithms hinges upon how economically this correlation pattern can be exploited. There are different approaches to proceed to this end:

- One can devise a sort of predictor, which predicts the future based on the past. In this case only the parameters of this predictor architecture and the quantized prediction error must be transmitted plus the initial part of the information process. Since quantization is a lossy business, this is a lossy scheme [1.4.3, 1.4.8].
- Analyzing the correlation structure of a random process, one can decide what are most important (or in other word, principal) components of the information which are worth to be preserved while the others can be dropped from the signal. These types of methods are rooted in principal component analysis and in transformation based methods [1.4.4].
- Observing relative frequencies of binary strings in a source file one can apply coding rooted in information theory to decrease redundancy [1.4.3].
- In order to get rid of repetition, one can set up a dictionary containing the most frequently occurring phrases in a text and then replace words by references to the dictionary (pointers and lengths...etc.) [1.4.9].

1.4.2. Algorithms

In this section we introduce different compression algorithms. The aim here is rather at communicating the underlying concept than to yield rigorous descriptions. For full algorithmic details see [1.4.6, 1.4.7]

Lossy schemes for data compression

In this section we describe algorithms which compresses analog data by reducing bandwidth at the expense of information loss.

Differential PCM

The samples of an information process (e.g. speech) exhibit strong correlation, which implies that the past samples can determine the present one. However, in this case it may suffice to encode only the difference of the present samples and its estimate obtained form the past samples which in turn results in a more economical use of bits. (As the past samples determine the present one statistically, this difference does not have much information and it can indeed be encoded by a smaller number of bits than it would be needed to encode the original sample. To pursue this line of reasoning even further, sometimes it is enough to indicate whether this difference is positive or negative using only one single bit, as will be seen in the case of Delta Modulation).

The coding principle which is developed to take advantage of the temporal correlation of the information process is officially termed as Differential PCM (D-PCM). The coder uses the predicted samples based on the AutoRegressive (AR) model given as $\hat{x}_n = \sum_{j=1}^{M} w_j x_{n-j}$. One can determine the optimal filter coefficients w_j, $j = 1,...,M$ by minimizing the mean square error given as $E\left(x_n - \hat{x}_n\right)^2$, yielding the following optimization task

$$\mathbf{w}_{opt} : \min_{\mathbf{w}} E\left(x_n - \sum_{j=1}^{M} w_j \xi_{n-j} \right)^2$$

which can be reduced to a set of linear equations (often termed as Yule-Walker equation)

$$\mathbf{R}\mathbf{w}_{opt} = \mathbf{r}\,, \qquad (1.4.1)$$

where $r_i = E(x_n x_{n-i})$ and $R_{ij} = E\left(x_{n-i} x_{n-j}\right)$.

There is an efficient method for a recursive solution developed by Levinson and Durbin [1.4.2].

The structure of DPCM coder is shown by the figure below

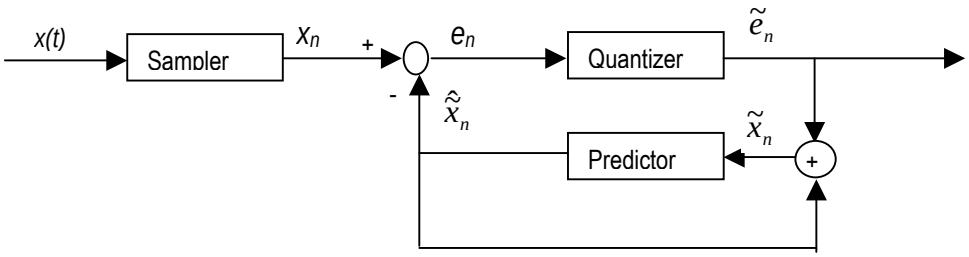


Figure 1.4.2. DPCM coder

The basic equations describing D-PCM coder/decoder scheme are given as follows:

$e_n = x_n - \hat{\tilde{x}}_n = x_n - \sum_{j=1}^{M} w_j \tilde{x}_{n-j}$, is the input to the quantizer, $\tilde{e}_n$ denotes the quantizated error and

$\hat{\tilde{x}}_n = \sum_{j=1}^{M} w_j \tilde{x}_{n-j}$.

(In the forthcoming discussion "tilde" always refers to the quantized samples. It is noteworthy that the real predictor operates with quantized samples, that is why the need for the combination of "hat" and "tilde".)

Adaptive DPCM (ADPCM)

So far we have assumed weakly stationary processes. In reality speech or video processes have time varying correlation.

The lack of stationarity entails a time varying SNR, even in the case of uniform quantization which depends on the current statistics of the input signal. To circumvent this problem, one can use an adaptive quantizer where the quantization level (originally denoted by Δ) is not constant but time varying as well, according to the rule $\Delta_{n+1} = \Delta_n M(n)$, where *M(n)* is a multiplicator depending on the quantized

value of the current sample. The possible values of *M(n)* optimized for speech signal were given by Jayant [1.4.6]. They are listed in the following table regarding 2-, 3- and 4-bit ADPCM:

Values of M(n)	2-bit quantizer	3-bit quantizer	4-bit quantizer
M(1)	0,80	0,90	0,90
M(2)	1,60	0,90	0,90
M(3)		1,25	0,90
M(4)		1,70	0,90
M(5)			1,20
M(6)			1,60
M(7)			2,00
M(8)			2,40

Another implication of the nonstationarity that equation (4.1) does not hold, as it was derived for stationary processes in the weak sense. Therefore not only the quantization levels, but also the predictor coefficients should be changed adaptively, by using the following set of equations:

$$w_l(k+1) = w_l(k) - \delta\left(x_n - \sum_{j=1}^{M} x_{n-j} \right) x_{n-l} \quad l = 0,...,M \ , \qquad (1.4.2)$$

where δ is the so-called "learning" coefficient. (Here index *n* refers to the time instant in the corresponding time sequence, whereas *k* indicates the current step of the learning algorithm.) This algorithm is called Robbins-Monroe stochastic approximation and the interested reader can find more details about its convergence in [1.4.4].

When using (1.4.2) the predictor coefficients must be transmitted to the receiver time-to-time which needs additional bandwidth. In order to get rid of this burden the receiver side has its own predictor which oerates on $\tilde{x}_n$ rather than x_n. Nevertheless, $\tilde{x}_n$ differs from x_n by the qunatization noise. Therefore, the convergence of such type of predictor is still an open question for further research.

ADPCM can not only be used for copressing speech but image siganl as well. Here one can take advantage of the twofold correlation of the images, when not only the neighbouring pixels but as well as the consecutive frames exhibit corrlation. Therefore, ADPCM can be used both for inter- and intraframe coding. The related algorithms are sismilar to the one derived for speech encoding.

Delta modulation

The simplest method of taking advantage of the temporal correlation of speech signal is referred to as Delta Modulation (DM). This architecture contains a two level (0,1) quantizer with a first order predictor. Since the coder acts as a discrete differentiator, the decoder contains an integrator to recover the signal as depicted as follows

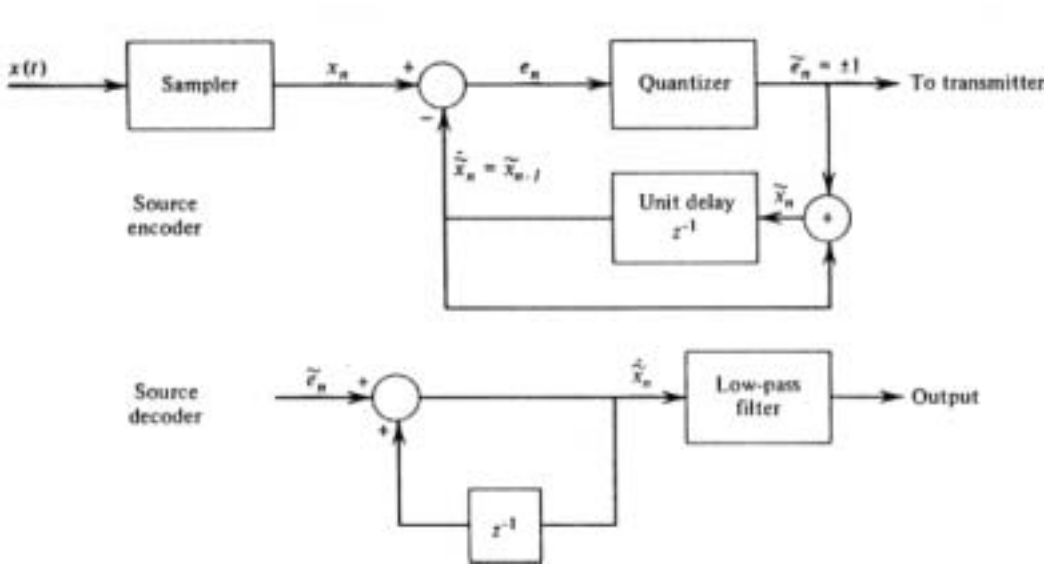


Figure 1.4.3. The coder and decoder of delta modulation system

The delta modulation algorithm is equivalent with approximating the original waveform by a staircase function. In order to have good approximation, the original waveform must vary slowly relative to the sampling rate. This requirement entails that the sampling rate must be at least 5 times higher than the Nyquist rate.

Using the delta modulator, two type of distortion can occur. The first one is referred to as granular noise which is due to the quantization step size Δ. The second one results from the finite "tracking" ability of the delta modulator and termed as "slope-over-load" distortion. The next figure demonstrate the effects of the two type of distortions.

In order to mitigate the effect of these distortions adaptive DM came into use in which the step size is varied adaptively subject to the following rule [1.4.8]:

$$\Delta_n = \Delta_{n-1} K^{\tilde{e}_n \tilde{e}_{n-1}}$$

where $K \geq 1$ constant is selected to minimize the distortion.

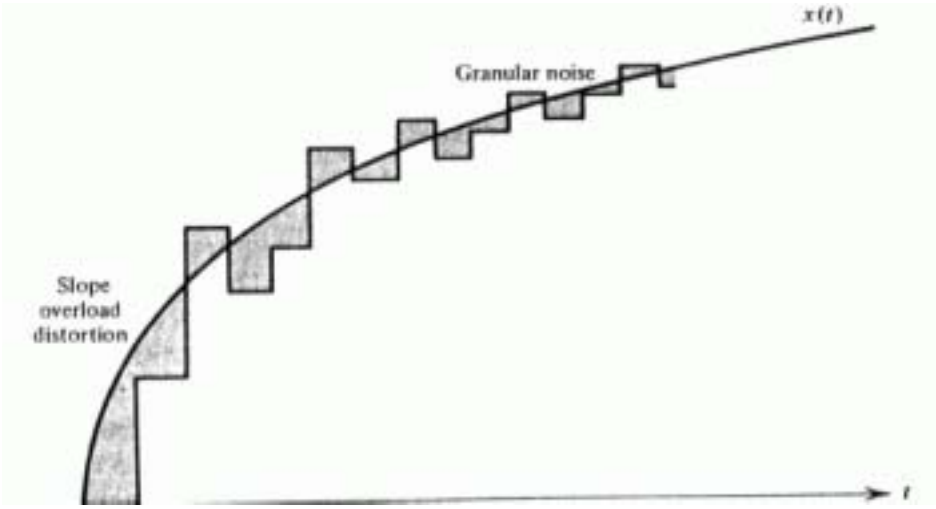


Figure 1.4.4. The example of slope overload and granular noise in delta modulation

There are several other adaptive Delta Modulator which have been published in the literature (for further reference see[1.4.8]). One of them is the so-called “continuously variable slope Delta Modulator” where the quantization levels are modified subject to the following rule: $\Delta_n = \alpha\Delta_{n-1} + \beta_1$ if $\tilde{e}_n$, $\tilde{e}_{n-1}$ and $\tilde{e}_{n-2}$ have the same sign, or $\Delta_n = \alpha\Delta_{n-1} + \beta_2$ otherwise.

Transform based methods for image compression

Transform based methods are also lossy schemes which discard part of the information. Of course the process of trading off information for a smaller bandwidth must be carried out in a structured manner. For example, not so many users would tolerate to discard a quarter of an image. In order to make intelligent guesses what to discard, one has to "prioritize" the data in some ways. Mathematically speaking, a new representation of the source information must be sought which reflects upon what is important and what is not in the data. Then we can proceed discarding the least important details until we achieve a prescribed quality limit (the amount of lost information does not exceed a given threshold). In this context decompression means carrying out the inverse transformation on the "truncated" representation of the original data. Since this method became highly popular in the case of image compression, in the forthcoming discussion we speak of images instead of general source information.

The general method of transform based compression is demonstrated by the next block diagram:

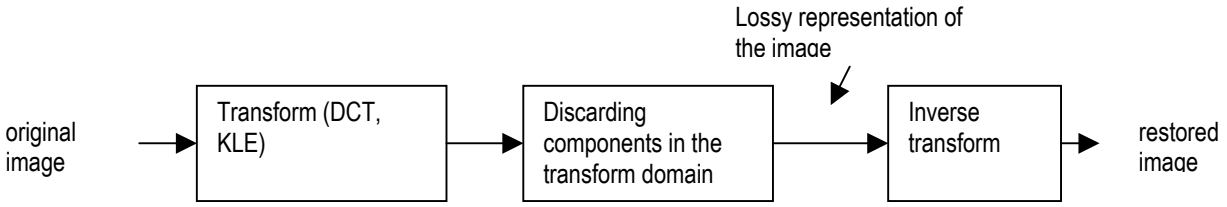


Figure 1.4.5 The principle of transform based compression

The mathematical transform to represent the image (on which basis one can decide what to discard) is either the Karhunen Loeve Expansion (KLE) or the Discrete Cosine Transformation (DCT).

Data compression by Principal Component Analysis (PCA)

This method is based on the Karhunen Loeve Expansion of the original image. More precisely, the image is perceived as an *N* dimensional random vector **x** (where **x** is assumed to be a column vector and $\mathbf{x}^{\mathbf{T}}$ is a row vector), the covariance matrix of which is $\mathbf{R} = E\left(\mathbf{x}\mathbf{x}^{\mathbf{T}}\right)$. (Note that $\mathbf{x}\mathbf{x}^{\mathbf{T}}$ denotes the outer product which results in a matrix and *N* denotes the number of pixels in the image.) It is known that the covariance matrix is Hermitic (symmetric in real case), therefore its eigenvalues are real and its eigenvectors are orthogonal to each other. The eigenvalues and eigenvectors are denoted as $\mathbf{R}\mathbf{s}^{(i)} = \lambda_i \mathbf{s}^{(i)}, i = 1,.., N$. It can also be shown that $\sum_{i=1}^{N} R_{ii} = \sum_{i=1}^{N} \lambda_i$. Since $\mathbf{s}^{(i)}, i = 1,..., N$ form an orthonormal basis, one can reconstruct **x** as $\mathbf{x} = \sum_{i=1}^{N} y_i \mathbf{s}^{(i)}$, where $y_i = \mathbf{s}^{(i)T}\mathbf{x}$

Compression takes place as follows:

Let us arrange the eigenvectors of the covariance matrix in a monotone decreasing order $\lambda_1 > \lambda_2 > ... > \lambda_M > ... > \lambda_N$. We consider the compressed image in the transform domain as a set of components $(y_1,..., y_M)$ from which the lossy

restoration of the image is given as $\tilde{\mathbf{x}} = \sum_{i=1}^{M} y_i \mathbf{s}^{(i)\mathrm{T}}$. Loss of information occurs, since we have discarded the components $(y_{M+1},...,y_N)$.

One can then prove (see [1.4.4]) that $E\left(\left\|\mathbf{x}-\tilde{\mathbf{x}}\right\|^2\right) = \sum_{i=M+1}^{N} \lambda_i$. Since the eigenvalues have been arranged in a monotone decreasing order the mean square error is minimal ($\sum_{i=M+1}^{N} \lambda_i$ is the smallest sum containing the last *N-M-1* eigenvalues). As a result PCA accomplishes the general objective elaborated in the introduction of this section, namely it represents the data in a transform domain, where one can prioritize what is important and what is not, and then discard the least significant details. "What-is-important" and "what-is-not" is mathematically defined by KLE. As a result, to compress images with a prescribed level of error ε, for which $E\left(\left\|\mathbf{x}-\tilde{\mathbf{x}}\right\|^2\right) \le \varepsilon$, one has to proceed as follows:

1. Determine the eigenvectors $S = \left\{\mathbf{s}^{(i)}, i = 1,...,N\right\}$ and eigenvalues $\Lambda = \left\{\lambda_i, i = 1,...,N\right\}$ of the covariance matrix $\mathbf{R}$.

2. Arrange the eigenvalues in a decreasing order $\lambda_1 > \lambda_2 > ... > \lambda_M > ... > \lambda_N$.

3. Determine *M* for which $\sum_{i=M+1}^{N} \lambda_i \le \varepsilon$.

4. Compress the image by calculating $y_i := \mathbf{x}^{\mathrm{T}} \mathbf{s}^{(i)}$ *i*=1,...,*M* (and send the vector $\mathbf{y} = (y_1,...,y_M)$ which is the "narrowband" equivalent of the image)

5. At the receiver side restore the lossy version of the original image as $\tilde{\mathbf{x}} = \sum_{i=1}^{M} y_i \mathbf{s}^{(i)\mathrm{T}}$

The flowchart of the PCA scheme is indicated by the following figure:

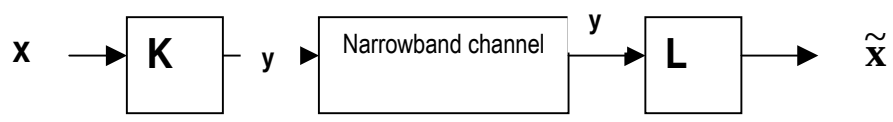


Figure 1.4.6 Data compression by PCA

Where **K** and **L** are matrices of type *MxN* and *NxM*, respectively and are defined as follows: $K_{ij} = s_j^{(i)}\ i = 1,...,M\ \ j = 1,...,N$ and $L_{ji} = s_j^{(i)}\ j = 1,...,N\ \ i = 1,...,M$.

Since the required transformations are linear, high speed image compression becomes possible based on modern high performance DSP systems (such as the TI 6x family).

However, there are some difficulties which can arise when one sets on implementing this algorithm:

- the correlation matrix of the image is not known but it is to be estimated on-line from the successive frames;
- PCA needs the calculation of the eigenvalues and the eigenvectors which amounts to performing a highly complex algorithm.

To circumvent this difficulties one can use adaptive PCAs, like "Kernel-based PCA" [1.4.4] or "Hebbian based PCA" introduced by Oja [1.4.4]. The latter method provides real-time and highly efficient compression algorithms which can be implemented by neural networks.

Image compression by Discrete Cosine Transform

In this case, we use a 2D Discrete Cosine Transformation (DCT) to represent the image and get rid of those spatial frequencies which will not fundamentally alter the quality of the image.

DCT is defined as follows:

$$\hat{f}(u,v) = \sum_{j,k=0}^{N-1} f(j,k) C(u) \cos\frac{(2j+1)u\pi}{2N} C(v) \cos\frac{(2k+1)v\pi}{2N}$$

where $f(j,k)$ represent the spatial information of the image (the value associated with pixel *(j,k)*) and $C(0) = 1/\sqrt{N}$ and $C(u) = 2/\sqrt{N}$ if $u \neq 0$.

The inverse transform is given as follows

$$f(i,j) = \sum_{j,k=0}^{N-1} \hat{f}(u,v) C(u) \cos\frac{(2j+1)u\pi}{2N} C(v) \cos\frac{(2k+1)v\pi}{2N}.$$

Data compression is achieved then by discarding more aggressively the "tail" of DCT asserting that high frequency spatial components add minor details to the image. JPEG coding is based on this principle and contains the following steps:

1. Process information locally. Divide the image into blocks containing 8x8 pixels.

2. Transform. On each block a DCT is performed to expose spatial properties.

3. Quantize. The coefficients are rounded off which reduces the amplitudes of the transformed coefficients. A more aggressive reduction is made on coefficients belonging to higher spatial frequencies.

4. Encode. Apply a lossless Huffman coding for further compression.

As a result, the decompressing algorithm contains the following blocks:

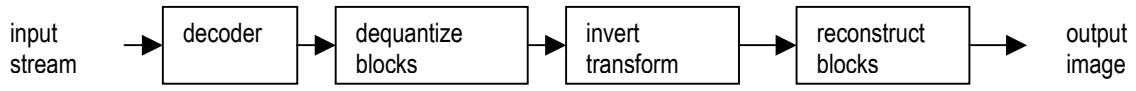


Figure 1.4.7 Data de-compression by JPEG

For further details on JPEG see [1.4.5].

Lossless algorithms for data compression

In the forthcoming discussion we summarize the basic algorithms used for lossless data compression. The main attribute of these methods that the original data can be fully recovered from the compressed data. For a full armory of these algorithm see [1.4.1, 1.4.2, 1.4.7]. There are two different approaches to lossless data compression:

- replacement schemes which are basically motivated by the statistical methods of information theory;
- dictionary based methods, which refer to the original text by using a set of pointers and length indicators.

Replacement schemes

Let us consider the binary data as a long file denoted by *W* while the compressed file is denoted by *U*. Lossless compression means that *W* is totally recoverable from *U.* Sometimes *W* is there available for excessive pondering on how to carry out the compression. Sometimes *W* is a high speed data stream "coming at the compressor" at a Mbit/sec rate. Therefore, the compression algorithm is not only qualified by the compression rate it achieves but by the speed of its operation, as well. There is not much use of a compression algorithm which needs almost as large complexity to restore the original file *W* as the length of *W* itself (here complexity is meant in terms of length of binary data).

Compression via replacement schemes is defined as follows:

Let us assume that a set of parsing-words $S = \{s_1,...,s_m\}$ is given. *W* is parsed into parsing-words if it is given as a sequence $W = (s_{i_1}, s_{i_2},..., s_{i_v})\ \ s_{i_j} \in S$. Furthermore, one defines an encoding scheme $C : S \to R$, where *R* represents the set of codewords $R = \{r_1,..., r_m\}$. If one replaces each s_{i_j} with a corresponding codeword r_{i_j}, then the compressed file $U = (r_{i_1}, r_{i_2},..., r_{i_v})\ \ r_{i_j} \in R$ is obtained. If the average length of the codewords r_i are smaller than the average length of the parsing words s_i, then length of *U* can be much smaller than *W*.

For example, let assume that $s_1 = 0\ \ s_2 = 10\ \ s_3 = 110\ \ s_4 = 1110\ \ s_5 = 11111$

$W = 11111011111110111010111101110110$ (30 bits long). As a result *W* can be parsed into the sequence $W = s_5 s_2 s_5 s_4 s_4 s_5 s_1 s_4 s_3$. Furthermore, let us choose the following encoding scheme: $r_1 = 1111\ \ r_2 = 1110\ \ r_3 = 110\ \ r_4 = 10\ \ r_5 = 0$. Then the resulting file is $U = 01110010100111110110$, which is only 20 bits long. The underlying ideas to carry out this compression are drawn from information theory. Namely, the following conditions should be met:

1. The codeset *R* should contain codewords which satisfy the prefix condition (easy and unique decoding from *U* to *W*).
2. The codeset must be constructed in such a manner that short codewords should be used to code those parsing words which frequently occur in the source file, while long codewords should be used to code those parsing words which occur less frequently in the text.

As a consequence, any traditional coding methods can be used for data compression if the underlying statistics can be estimated "on-the-fly". More precisely, adaptive Shannon-Fano or Huffman coding can achieve relatively large data compression rate. These schemes basically come down to the question of how to obtain the underlying statistics. The simplest method is to calculate relative frequencies on the source file *W*. In this case, one can obtain the following statistics:

$$f_i = \frac{\text{number of occurance of } s_i \text{ in the source file}}{\text{total number of parsing words in the source file}}$$

In our example $f_1 = \frac{1}{9}\ \ f_2 = \frac{1}{9}\ \ f_3 = \frac{1}{9}\ f_4 = \frac{3}{9}\ \ f_5 = \frac{3}{9}$. The problem with this scheme is that in order to obtain an accurate estimation of the probabilities, a

relatively long file must be observed. However, when the file arrives as a Mbit/sec datastream then one is better off estimating this relative frequencies recursively.

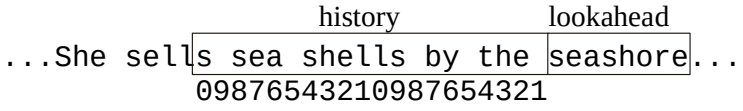
Of course when adaptive Huffman encoding takes place then certain measures should be introduced to maintain the tree of the Huffman encoding. The corresponding methods were originated by Gallager and Knuth. The interested reader can find further details in [1.4.3].

Dictionary based methods

Dictionary based methods use a list of phrases ("the dictionary") and replace fragments of the source text by pointers to the list. Compression occurs if the pointers require less space than the corresponding fragments. In addition to the pointers the hand-over of the dictionary must also be considered. Since we want to perform data compression in communication systems, we are concerned with creating the dictionary and setting up the pointer on-the-fly.

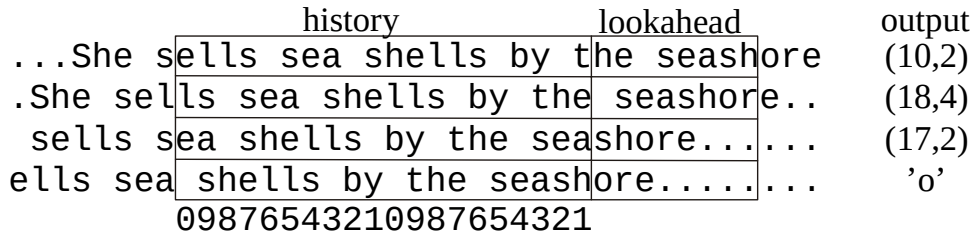
Adaptive dictionary methods date back to the seventies when Ziv and Lempel published their landmark algorithm. The corresponding schemes are known as LZ77 and LZ78, respectively. Applications based on LZ77 are LHarc, PKZIP, GNU zip, Info-Zip , whereas LZ78 is used in MODEM communication and in GIF graphics format. To illustrate the underlying idea, here we summarize the LZ77 type of sliding window schemes.

In this case a two-side window slides through the source text, the first side is termed as *lookahead*, whereas the second one is the *history*. For example



In the simplest case, the history and lookahead are of fixed length and the dictionary contains all phrases which start in the history and are not longer than the lookahead. The ideas is to replace the initial segment of the look-ahead with a pointer into the dictionary and then move the window ahead. In the example given below, the starting segment of 'he' matches the second two charters of the dictionary word 'shell' (at the position of 10 in the dictionary). Therefore, the compression algorithm would yield a triple as an output *(coordinate, length, character)*. The third component is the next character from the source. According to the example it is

(10,2,_), where _ is the space character. The *character* component allows to proceed even if there is no match in the history. The following figure shows a few steps of LZ77,



while the next table indicates the parameters of practical implementation

Length of history	Length of look-ahead	Coordinate	Length
4096 bytes	4096 bytes	12 bit	4 bit

A single control bit is used to differentiate *(coordinate, length*) pairs from characters. Therefore the cost of transmitting a *(coordinate, length)* pair is 17 bit. Therefore, only those source fragments will be replaced by a pointer which have length longer than 2 bytes.

The next table shows the some typical compression rates achieved by the LZ77 algorithm on the "Calgary corpus" (a set of files on which compression efficiency is frequently tested).

File	Kbyte	Compression rate achieved by LZ 77
bib	109	43 %
book1	751	51.8 %
geo	100	80.3 %
obj1	21	57.2 %
pic	501	20.9 %
progc	39	41.7 %
Average		49.2 %

In the case of LZ78 the source text is parsed into phrases, and a new phrase enters into the dictionary by adding a single character to an existing phrase. This method is used in UNIX Compress, V.42*bis* MODEM communication and the GIF graphics format.

In the case of Voice over IP the G.&x family is applied which in certain cases can achieve 6.4 Kbps speed, however it needs high computational power.

Other techniques, such as vector quantization can also prove to yield efficient compression (see [1.4.4]).

1.4.3. Standards & patents

As was mentioned before speech is traditionally compressed by using adaptive predictive method and the corresponding system with given parameters is referred to as ADPCM. In practice 4 bit ADPCM system is used the most frequently which are appropriate for applications at 32kbps and 24kbps in contrast with the traditional 64 kbps. The 32 kbps speech compression is fully spelt out in the standard ITU G.726, which proposes the use of a 4-bit quantizer.

For mobile telephony, UNIX based implementation of GSM 06.10 full rate transcoding are available throughout Europe. resulting in a 13 Kbps bandwidth using long term prediction combined with residual pulse excitation.

Speaking of compression of music one has to mention the standard Music Instrument Digital Interface (MIDI), which is adopted by almost the entire music industry. A MIDI message conveys musically significant events, such as a key being pressed ...etc. Every instrument has a MIDI code assigned to this message (e.g. the piano is 0 violin is 40). There are 127 instruments listed. In the case of MIDI system the computer receives messages and generates music based upon them. The synthesizer understands all 127 instruments and generate a spectrum instructed by the received MIDI codes. The advantage of MIDI is the enormous reduction in bandwidth. Its drawback that it needs a synthesizer at the receiver side and each one may generate a slightly different musical quality.

The best known standard to compress still images is JPEG (Joint Photographic Expert Group). JPEG encodes 24-bit RGB video of size 640x480 pixels. It first calculates luminescence and chrominance information from the RGB image and it compresses chrominance by a factor two as human vision is not sensitive to colour information. Then the luminance and chrominance matrix is divided into 8x8 blocks and a DCT is performed on each block. Then the less important coefficients of DCT are eliminated block by block by using a quantization algorithm. In this way we have got rid of the higher spatial frequency components. Then run length encoding is carried out followed by a Huffman encoding. JPEG

standard appears to be very complicated but it is rewarding in the sense of being capable of achieving 20:1 compression rate.

To extend compression to moving pictures the MPEG (Motion Picture Experts Group) standard came into use. The first one (MPEG-1) was developed to reduce NTSC video to 1.2Mbps bandwidth. It is also used for storing movies in CD ROM in CD-I and CD-Video format. Then MPEG-2 followed to compress broadcast video (NTSC and PAL) to the range of 4-6Mbps. It was even expanded to higher resolution HDTV applications. MPEG-4 is for medium resolution video conferencing. In MPEG, audio and video are compressed in synchron by using 90 kHz system clock. Audio is compressed by using either 32kHz, 44.1KHz or 48Khz. It applies FFT, nonlinear quantization, Huffman coding and other sophisticated methods. Regarding video, MPEG gets rid of spatial redundancy by using the JPEG standard. However, successive frames do not differ too much on a short time scale, therefore MPEG also gets rid of time redundancy. This is mainly done by using block by block difference with the last frame and bi-directional difference with the last and next frame.

Concerning dictionary based methods GIF graphics format is based on the LZ78 algorithm (Lempel, Ziv, Welch). This algorithm has also become infamed because of legal fights regarding patent issues. The companies which adopted this algorithm are IBM, Unysis. The basic controversy is around the opinion of the patent offices, which maintain that algorithms are not patentable, but an algorithm used to solve a particular problem qualifies for a patent. This indeed is a controversy to be tackled.

Summary

In this chapter different data compression algorithms were described. Lossy schemes were basically used to compress speech and video, while lossless schemes were developed to compress data. The corresponding ideas were addressed in the structure of discussing algorithms, standards and patents.

One must also mention the considerable research conducted in the field of wavelets and fractals to use them for image compression. Nevertheless, due to the limited size of the chapter, the descriptions of the related algorithms are beyond the scope of the present treatment.

Available websites on data compression and related materials

On fractal image coding: http://inls.ucsd.edu/Research/Fisher/Fractals

On GNU projects: http://www.gnu.ai.mit.edu

On Info-Zip: http://www.cdrom.com/pub/infozip/

On JPEG: ftp://ftp.uu.net/graphics/jpeg

On wavelet based compression: http://www.cs.dramouth.edu/~gdavis.

On PNG and LZ77: http://www.wco.com/~png/

References:

[1.4.1] Bell, T., Witten, J.: *Text compression*, Prentice-Hall, 1990

[1.4.2] Durbin, J.: "Efficient estimation of parameters in moving average models", *Biometrika*, Vol. 46, 1959

[1.4.3] Hankerson, D., Harris, A., Johnson, P.*: Introduction to information theory and data compression*, CRC Press, 1998

[1.4.4] Haykin, S.: *Neural networks - a comprehensive foundation*, Prentice-Hall, 1999.

[1.4.5] Independent JPEG Group, JPEG SW Release 6. ftp://ftp.uu.net/graphics/jpeg

[1.4.6] Jayant, N., Noll, P.*: Digital coding of waveforms, principles and applications to speech and video*, Prentice-Hall, 1984

[1.4.7] Nelson, M., Gailly, J.: *The data compression book*, M&T Books, 1996

[1.4.8] Proakis, J.: *Digital communications*, McGraw-Hill, 1983

[1.4.9] Ziv, J., Lempel, A.: "A universal algorithm for sequential data compression", *IEEE Trans. on Information Theory*, May, 1977, pp. 337-343

1.5. Modulation methods

István Frigyes dr., author

György Dallos dr., reviewer

1.5.1. Introduction

Information while transmitted, stored and processed is present in the form of *electrical* or *electromagnetic* signals. This signal is an electrical voltage, current, electromagnetic field strength, optical power or else. In some cases the source produces *electrical* signals, e.g. data sources, ECGs (Electrocardiographs), EEGs (Electroenkephalographs) and others. In other cases the signal to be transmitted, stored or processed appears in some form to be perceived for human sense, such as voice or sight. In these cases some transducer forms the electrical signal.

Whatever the physical appearance of the signal is it can be modeled as a *real time function*. Properties, mathematical forms of these time functions are dealt with in section 1.1. In what follows, in most cases real physical signals and abstract signals modeling these mathematically will not be distinguished.

Electrical source signals or signals transuded in electrical form as well as their digitized versions are usually called *baseband* signals. Fourier transform of these contains often zero frequency or even if not their lowest frequency component is rather low, without specifying more precisely what is meant under high or low.

Signals are processed, stored usually or at least very often in their baseband form. Sometimes they are also transmitted in baseband. However, very often it is more favorable moreover only possible to transmit signals, both wirebound and wireless, in a form transduced to a somewhat or much higher frequency band. The operation transducing frequency band and its inverse are called modulation and demodulation, respectively. The present chapter deals with these operations and with concepts related to these.

1.5.2. Basic concepts

Let the real time function containing the information be *s*(*t*). This can be a known (i.e. a deterministic) time function or a sample functon of a random process. Let further a sinusoidal signal v(*t*) be

$$v(t) = \sqrt{2}A\cos(\omega_c t + \Phi) \qquad (1.5.1)$$

with *A* the rms amplitude of the signal given in (Watt$^{1/2}$)

$\omega_c = 2\pi f_c$ angular frequency (in rad/sec subscript *c* stands for *carrier*

Φ the initial phase.

Modulation defined in the Introduction is usually achieved by *inscribing* signal *s*(*t*) somehow into the amplitude or into the argument of the sinusoidal signal – i.e. by modulating it. Modulated parameter(s) become(s) in this way time functions being in *unique relationship* with *s*(*t*). Thus it becomes possible to demodulate this modulated signal, forming the inverse of this function.

From the point of view of the sinusoidal signal the general form of a modulated signal is

$$x(t) = \sqrt{2}Am(t)\cos\left[\omega_c t + \Phi + \vartheta(t)\right] \qquad (1.5.2)$$
$$m(t) = f\left[s(t)\right];\ \vartheta(t) = g\left[s(t)\right]$$

where m(t) is the amplitude modulation and ϑ(t) is the angle modulation.

So called *quadrature* representation

$$x(t) = Aa(t)\cos(\omega_c t + \Phi) - Aq(t)\sin(\omega_c t + \Phi) \qquad (1.5.3)$$

is equivalent to the above *modulation representation*. Here

a(*t*) is the in-phase component and *q*(*t*) is the quadrature component of the modulated signal.

Relationship between these two pairs of functions is

$$a(t) = m(t)\cos\vartheta(t);\ q(t) = m(t)\sin\vartheta(t) \qquad (1.5.4)$$
$$m(t) = \sqrt{\left[a(t)\right]^2 + \left[q(t)\right]^2};\ \mathrm{tg}\,\vartheta(t) = \frac{q(t)}{a(t)}$$

In some cases one of these representations, in other cases the other one is more advantageous.

In some cases carrier is a non-sinusoidal periodic function. Its form in this case is

$$v(t) = \sum_{k=-\infty}^{\infty} Ap(t - kT + \tau_0); A = \frac{1}{T}\int_0^T [p(t)]^2 dt \qquad (1.5.1a)$$

Support of $p(t)$ is $(0,T)$; it has to fulfill relatively simple constraints. These are about the same as needed for the existence of a Fourier series of $p(t)$. We'll not deal with this version in great detail. What were said or will be said about the sinusoidal carrier are valid – by changing the variables – for this case as well.

To conclude this section it is mentioned that the Fourier transform of signals will often be mentioned. This wording assumes a deterministic signal. Our statements are in most cases valid also for the spectral density function of random processes; the converse of this is not always right: what is stated about spectral density is not necessarily valid for the Fourier transform of deterministic signals.

1.5.3. The analytic signal and the complex envelope

The complex envelope concept being applicable for describing of modulated signals is not only of basic importance in the investigation of these signals – we'll also use it in what follows – but also in the description of propagation phenomena, of noise, it is *sine qua non* in the computer simulation of communication systems, of digital signal processing etc. Therefore we'll deal with it in some detail.

Quadrature form of a modulated signal can be written in this way also:

$$x(t) = A \cdot \mathrm{Re}\left\{[a(t) + jq(t)]e^{j\omega_c t}\right\} \qquad (1.5.5)$$

where Re means "Real part of" and the initial phase is omitted being insignificant.

Complex quantity *a(t)+jq(t)* is the *complex envelope* of the signal. In the sequel we'll deal with the general concept of the complex envelope as well as with that of *narrowband signals*.

It is known that Fourier transform $S(\omega)$ of a real function is conjugate symmetric (i.e. $S(-\omega) = S^*(\omega)$). Thus $S(\omega)$, $\omega > 0$ determines the signal uniquely. We

know that for getting the Fourier transform we must add to it the appropriate part at ω < 0. So the signal can be characterized by $\hat{S}(\omega)$ instead of $S(\omega)$ where, by definition

$$\begin{aligned}\hat{S}(\omega) &= S(\omega) + \operatorname{sign}(\omega)\cdot S(\omega) = \\ &= S(\omega) + j[-j\cdot\operatorname{sign}(\omega)\cdot S(\omega)]\end{aligned} \qquad (1.5.6)$$

Note that for $\omega>0$ $\hat{S}(\omega)$ = 2S(ω); for $\omega<0$ $\hat{S}(\omega)$ = 0.

Of course, the inverse Fourier-transform of $\hat{S}(\omega)$ can also be calculated. For that note that the second term is a product of two factors. Thus in the inverse Fourier transform their convolution must be formed:

$$\mathbf{F}^{-1}[-j\cdot\operatorname{sign}(\omega)\cdot S(\omega)] = s(t) * \mathbf{F}^{-1}[-j\cdot\operatorname{sign}(\omega)] \qquad (1.5.7)$$

The second factor is 1/t. Thus

$$\mathrm{F}^{-1}[-j.\operatorname{sign}(\omega).S(\omega)] = \int_{-\infty}^{\infty}\frac{s(\tau)}{t-\tau}d\tau = \hat{s}(t), \qquad (1.5.8)$$

i.e. the Hilbert-transform of So

$$\hat{s}(t) = \mathrm{F}^{-1}[\hat{S}(\omega)] = s(t) + j\cdot\hat{s}(t) \qquad (1.5.9)$$

Complex time function $\hat{s}(t)$ is called the *analytic function* of *s*(*t*).

The reason of this nomination is: if instead of *t* the complex variable *z* = *t* + *j.u* is applied, $\hat{s}(z)$ fulfills partial differential equations of Cauchy and Riemann, i.e. it is analytic. Real part of it is our time function *s*(*t*). Note that Fourier transform of the analytic function is zero for negative frequencies.

The analytic signal is important in dealing with modulated signals. Namely the analytic function of cos $\omega_c t$ is $e^{j\omega_c t}$. If the amplitude of this carrier is modulated by some band limited signal *a*(*t*) we get $a(t)e^{j\omega_c t}$. This is analytic if the bandwidth of *a*(*t*) is less than $f_C=\omega_C/2\pi$. Similarly, the analytic signal of $-j.\sin(\omega_C t)$ is also $e^{j\omega_c t}$. If modulated by a similar band limited signal *q*(*t*), it is still analytic. Thus the modulated signal can really be represented as

$$\operatorname{Re}\{[a(t)+jq(t)]e^{j\omega_c t}\} \qquad (1.5.10)$$

as long as the bandwidths of the modulating signals is less than the carrier frequency. This type of signals is called *narrow band*. (Note that the condition is *less* and not *much less*.) Signal $\tilde{s}(t) = a(t) + jq(t)$ is the *complex envelope* of the modulated signal.

(A counter example in brackets: if the bandwidth B>f_c, $[a(t) + jq(t)]e^{j\omega_c t}$ is *not* analytic and the signal is not its real part.)

Band limited modulating signal results nearly always in a narrow band signal. So in calculations it is sufficient to deal with complex envelope *a(t) + jq(t)*. We know then that its real part is the cosinusoidal part of the modulated signal and its imaginary part is its sinusoidal part. The complex envelope is a complex valued baseband signal; its Fourier transform can be got by shifting the transform of the analytic signal by -ω_c. That is there exists the following relationship between the signal shape and the complex envelope:

$$\tilde{x}(t) = [x(t) + j\hat{x}(t)]e^{-j\omega_c t}; x(t) = \mathrm{Re}[\tilde{x}(t)e^{j\omega_c t}] \qquad (1.5.11)$$

For computing the complex envelope we must form the analytic signal and multiply it by exp[-$j\omega_c t$]; conversely: multiplying the complex envelope by exp[$j\omega_c t$] and forming the real part we get the real time function.

In a similar way the equivalent lowpass filter $\tilde{H}(\omega)$ of a bandpass filter. Transfer function of the former is the -ω_c shifted version of that of the latter. And the expression of a signal transferred through the bandpass is

$$\begin{aligned} &\tilde{Y}(\omega) = \tilde{X}(\omega) \cdot \tilde{H}(\omega), \text{ resprctively} \\ &\tilde{y}(t) = \tilde{x}(t) * \tilde{h}(t) \end{aligned} \qquad (1.5.12)$$

An important characteristic is as follows. If the cosinusoidal component is only modulated $\tilde{X}(\omega)$ is, of course, conjugate symmetric. But $\tilde{H}(\omega)$ is or is not (a bandpass filter is or is not conjugate symmetric around the carrier frequency). If the latter case holds $\tilde{y}(t)$ is not real. This means that there is a crosstalk between the sinusoidal and the cosinusoidal carriers. There is no crosstalk if $\tilde{H}(\omega)$ is conjugate symmetric around the carrier frequency.

1.5.4. Analog modulation systems – amplitude modulation

As discussed to more detail in Section 1.1 an analog signal is represented by a time function being whatever, within some specification. This specification can be the support of the time function or of its Fourier transform, its power, its dynamic range or something the like. (For our younger readers it is worthwhile to tell: this nomination stems from the early years of computer techniques; then *analog* computers did exist besides of digital computers. In these to solve a mathematical problem an electric circuit *analog* to it was built – analog in the sense that it fulfilled the same differential equation, integral equation etc. – and the signal shapes were investigated. This is the basis of categorizing signals as analog or digital. A signal of continuous range and not precisely specified form is called analog.) Study of analog modulation systems deals with the problem: how to modulate a (sinusoidal) carrier by an analog signal.

Definitions and spectral characteristics

i. Double sideband non-suppressed-carrier (AM-DSB-NSC) In AM function $\vartheta(t)$ of Eq. (1.5.2) is zero and $m(t)$ depends linearly on $s(t)$. Thus the general form of an amplitude-modulated sinusoid is

$$x(t) = \sqrt{2}A[1 + h \cdot s(t)]\cos \omega_c t \qquad (1.5.13)$$

where h is called the modulation index.

Let us examine the Fourier transform of an AM signal (Figure. 1.5.1). Section 1.5.3 shows that the complex envelope of this signal is $[1 + h \cdot s(t)]$ and its Fourier transform is

$$\mathbf{F}[1 + hs(t)] = \delta(\omega) + hS(\omega) \qquad (1.5.14)$$

Thus theFourier transform of the analytic signal is

$$\hat{X}(\omega) = \sqrt{2}A[\delta(\omega - \omega_c) + hS(\omega - \omega_c)] \qquad (1.5.15)$$

while that of the real modulated time function

$$X(\omega) = \frac{\sqrt{2}}{2} A[\delta(\omega - \omega_c) + h.S(\omega - \omega_c) + \delta(\omega + \omega_c) + h.S(\omega + \omega_c)] \qquad (1.5.16)$$

Figure 1.5.1 Fourier transform of a modulating signal and of the analytic signal of the AM-DSB-NSC modulated signal

Note that the signal of Eq. (1.5.13) contains an unmodulated carrier. This means that its Fourier transform contains a spectral line at the carrier frequency; intensity of this line does not depend on the modulating signal. Further note that $X(\omega)$ contains $S(\omega)$ in an unchanged, with other words in an undistorted form; therefore AM is often called a linear modulation. And further: as $s(t)$ is a real time function, $S(\omega)$ is a conjugate symmetric function. Consequently $X(\omega)$ contains two side bands which are conjugate symmetric around the carrier frequency. thus we see that if the support of $S(\omega)$ is of width W the occupied bandwidth of the modulated signal is $2W$.

According to all that amplitude modulation as given in Eq. (1.5.13) is called *double sideband non-suppressed-carrier amplitude modulation*, abbreviated as AM-DSB-NSC. This type of modulation is wasteful in two respects. First in frequency as the whole information is contained in a band W and still we occupy two-times as much. It is also wasteful in power: the power of the unmodulated carrier does not contribute to the quality of the transmitted signal.

ii. Double-sideband suppressed carrier AM (AM-DSB-SC) More economic methods can be desirable. $x(t)$ of (1.5.13) can be modified not to have "wasted power". Then

$$x(t) = \sqrt{2}As(t)\cos\omega_c t \qquad (1.5.17)$$

In this case the carrier is simply multiplied by the modulating signal. Spectrum of the analytic signal is the most clear among those given above. It is

$$\hat{X}(\omega) = \sqrt{2}AS(\omega - \omega_c) \qquad (1.5.18)$$

iii. Single sideband AM (AM-SSB) As one sideband contains the total information to be transmitted it is reasonable, for economic frequency band application, not to occupy the second one. We get such a single sideband AM if we modulate the carrier by the analytic signal rather than by the signal itself. (Namely, as we know, in this case $s(t)$ has no negative frequency component and so $x(t)$ will have only one sideband.) I.e. in this case

$$x(t) = \sqrt{2}A.\left[s(t) \pm j\hat{s}(t)\right]\cos\omega_c t \qquad (1.5.19)$$

and we get the upper sideband if the positive sign and the lower one if the negative sign is taken.

Looking at Eq. (1.5.6) we see that to get $\hat{s}(t)$ we have to transmit *s*(*t*) over a filter of constant attenuation and phase shift of $\pi/2$. To realize such a filter is not very easy. (Of course an SSB signal can also be produced if we filter one of the sidebands; this can be done if the signal has no low-frequency components, e.g. the voice signal.)

iv. Vestigial sideband AM In lack of space we don't discuss only mention that practical difficulties of producing and demodulating SSB can be reduced if besides of one sideband we transmit (even a small) part of the other.

Demodulating AM signals

AM-DSB-NSC signals can simply be demodulated with an envelope detector. The price of this simplicity is paid in wasting power and frequency band. Note that envelope detection assumes that the instantaneous amplitude of the signal is positive; thus *h.s*(*t*)<1.

To demodulate a suppressed-carrier AM (single or double sideband) we must know the *phase* of the carrier. Thus a *reference signal* is needed, being at least in principle coherent with the received signal. Having this available a product demodulator can be applied. E.g. AM-DSB-SC:

$$s_d(t) = \sqrt{2}Ahs(t)\cos\omega_c t \times \sqrt{2}R\cos\omega_c t = ARhs(t)(1+\cos 2\omega_c t) \qquad (1.5.20)$$

with *R* the rms. amplitude of the reference signal. Double frequency term can easily be filtered.

It can be shown that if there is a phase difference δ between the locally generated reference signal and the received one (i.e. coherence is not perfect)

- with DSB-SC the demodulated signal is proportional to cos δ; thus frequencies must be *exactly* equal, otherwise cos δ is permanently changing between ± 1.
- In SSB although signal shape is essentially distorted power (i.e. absolute squared of the complex envelope) remains unchanged. The distortion is thus not perceived e.g. in voice transmission.

1.5.5. Analog modulation systems – angle modulation

We don't deal with angle modulation in details; we give a few definitions and spectral properties for a few cases.

Definitions

In angle modulation $m(t)$ = 1 in Eq. (1.5.2) and $g[]$ is in practice a linear operator. In practical applications there are three different g operators and corresponding $\vartheta(t)$ functions.

i. Phase modulation (PM)

$$\vartheta(t) = 2\pi h s(t); x(t) = \sqrt{2}A\cos\left[\omega_c t + 2\pi h s(t)\right] \qquad (1.5.21)$$

In this case carrier phase is proportional to the modulating signal; h is the modulation index.

ii. Frequency modulation (FM)

$$\vartheta(t) = 2\pi.\Delta F.\int^{t} s(t)dt; x(t) = \sqrt{2}A\cos\left[\omega_c t + 2\pi.\Delta F.\int^{t} s(t)dt\right] \qquad (1.5.22)$$

Instantaneous frequency is then proportional to the modulating signal. ΔF is frequency deviation.

iii. FM with preemphasis

Low frequency components of the modulating signal are suppressed, high frequency components enhanced and carrier frequency is modulated by this distorted signal. (This can improve SNR of the demodulated signal; spectral density if FM noise is proportional to ω^2.) In this case

$$\vartheta(t) = 2\pi.\Delta F.\int^{t} p(t) * s(t)dt; x(t) = \sqrt{2}A\cos\left[\omega_c t + 2\pi.\Delta F.\int^{t} p(t) * s(t)dt\right]$$

(1.5.23)

where $p(t)$ is the impulse response of the spectrum forming filter.

Spectral characteristics – some examples

In contrast to AM, FM transforms $s(t)$ into $x(t)$ in a nonlinear way. So the spectrum of $x(t)$ can not be written directly. In what follows a few special cases will be investigated.

i. Frequency modulation by one sinusoid

In this case $s(t)=\cos \omega_m t$ and $\vartheta(t) = \Delta F/f_m.\sin \omega_m t$. Modulated signal is periodic and so it can be by its Fourier series. Applying the Fourier series of cos(sin x) we get for the Fourier transform of $x(t)$

$$X(\omega)=\frac{A}{\pi\sqrt{8}}\sum_{n=-\infty}^{\infty} \mathrm{J}_n\left(\frac{\Delta F}{f_m}\right)\left[\delta(\omega-\omega_c-n\omega_m)+\delta(\omega+\omega_c+n\omega_m)\right] \qquad (1.5.24)$$

where J_n is the Bessel function of order n.

As seen it contains an infinite number of spectral lines the magnitude of which depends in a complicated way on the frequency deviation. If $\Delta F/f_m$ is very low there are only two spectral lines of non-negligible amplitude; if it coincides with the zero of J_0 there is no carrier component. etc.

ii. Narrow band FM, any signal

If the frequency deviation is much lower than the upper frequency limit of the modulating signal the modulation is similar to a linear one: shape of the two sidebands is equal to that of $\vartheta(t)$.

iii. Wideband FM, random modulating signal

Rather interesting case: there are two sidebands the shape of which is that of the probability density function of the modulating signal.

iv. Occupied bandwidth

As seen in (1.5.23) the occupied bandwidth is theoretically infinite. The practically occupied bandwidth is rather well described by the formula of Carson:

$$B=2(\Delta f+f_{\max}) \qquad (1.5.24a)$$

Demodulating angle modulated signals

In frequency modulation the instantaneous frequency of the modulated signal is proportional to the modulating signal. Thus for demodulation a frequency detector

is needed. Amplitude of its output is proportional to the frequency of the input. This requirement involves of course that the output signal must be independent from the input amplitude.

A mistuned resonator can be used to convert frequency variation into amplitude variation – see resonance curve of Figure. 1.5.2. (Abscissa noted is carrier frequency. It should coincide with the inflexion point of the resonance curve.) Output signal can then be detected by an envelope detector.

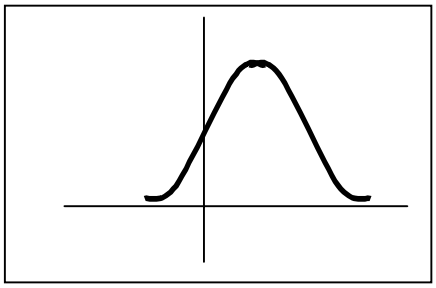


Figure 1.5.2. A mistuned resonator as frequency-to-amplitude converter

This frequency-to-amplitude conversion results, of course, in nonlinear distortion, usually unacceptably high. Applying two mistuned resonators and subtracting their detected signals can improve linearity. Characteristic of such a so-called frequency discriminator is shown in Figure. 1.5.3.

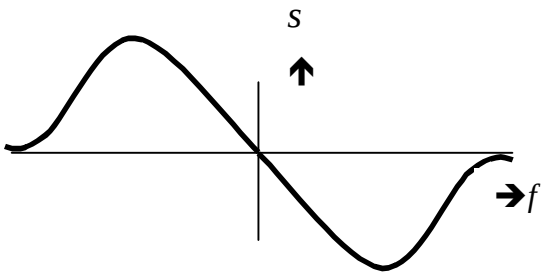


Figure 1.5.3 Characteristic of a frequency discriminator

To insure independence of the output signal from the input amplitude a limiter is used. So a complete frequency demodulator is composed of a frequency discriminator preceded by a limiter.

FM can also be demodulated by a PLL (Phase Locked Loop), see Figure. 1.5.4. In equilibrium of the PLL VCO frequency must *exactly* be equal to that of the input signal. Thus if the latter is varying control signal of the VCO must follow the

frequency variation – i.e. it yields detected frequency. Further modulation characteristic does not depend on the amplitude of the PLL input signal, thus PLL is also an ideal limiter.

Without discussing in detail it is mentioned that a PLL can also be used as an analog phase demodulator, also shown in the figure.

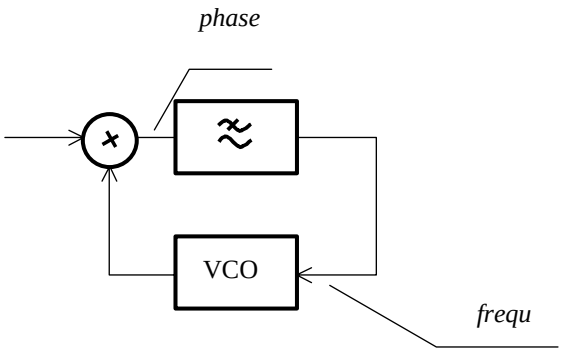


Figure 1.5.4 PLL as analog frequency and phase demodulator

FM in additive noise

Noise, not being taken into account yet modifies demodulated FM signal: it is also noisy. Without derivation the formula for signal-to-noise ratio is given below, and an important phenomenon i.e. threshold is briefly discussed.

Signal-to-noise ratio of demodulated FM signal corrupted by noise is given as

$$S/N = \frac{C}{N}\left(\frac{\Delta F}{f_{\max}}\right)^2 \qquad (1.5.24.b)$$

where *C/N* is the carrier-to-noise ratio in the receiver. (For this note that RF bandwidth is approximately given in Eq. (1.5.24.a).) Eq. (1.5.24.b) shows that by increasing frequency deviation S/N can be increased without increasing received power, (and, at the first glance without any limit). This is the main advantage of the application of FM.

However, validity of Eq. (1.5.24.b) depends on two conditions. i. Bandwidth of the envelope detector must not be higher than $f_{\max}$ – a condition easy to fulfill. ii. *C/N* must be rather high, higher than 7-10 dB or so. If it is lower, due to the threshold phenomenon already mentioned, discriminator doesn't work any more as an ideal frequency detector and *S/N* is lower than that of (1.5.24.b).

To deal with this problem take Figure. 1.5.5 into account. At high *C/N* signal+noise is with high probability close to the signal. Say it is encircling it, as shown in the figure. Phase is changing according to that (limits of the phase change are also shown in the figure). Derivative of the phase variation is frequency noise, being relatively low. Eq (1.5.24.b) was derived under this assumption. If, however, signal is low the same realization of the noise process as before causes a whole cycle slip of the phase. Derivative of that is a large noise pulse, so in this casethe received signal will be corrupted by impulsive plus Gaussian noise, significantly decreasing S/N.

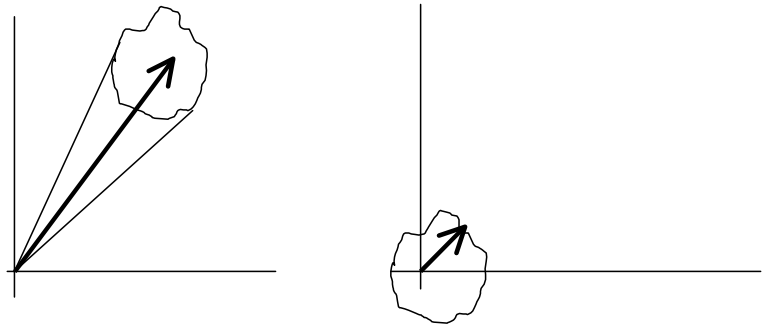


Figure 1.5.5. Phase variation of the noisy signal in the case of high value and of low value of C/N

In principle there is a similar phenomenon in the PLL-demodulator also, however, with significant quantitative difference. I.e. PLL bandwidth must be in the order of f_{max} rather then being in the order of $2\Delta F+f_{max}$. Thus cycle slips have the same probability at significantly lower input power than in the frequency discriminator. PLL is thus a very good means to achieve low threshold in an FM demodulator.

1.5.6. Digital modulation methods

While an analog signal is defined as one member of a rather wide class of functions in dealing with *digital* signals – transmitting, storing or processing them – one element of a finite-element signal set is dealt with. (Thus choice of modulation scheme means defining these signal shapes.) Number of elements of the signal set is *M*, each element is characterized by its *finite duration* (T or T_S), and by its *finite energy*. (The energy of a signal is defined as its square integral.)

Thus a digital signal stream is given as

$$x(t) = \sum_{k=-\infty}^{\infty} s_i(t - kT_s); i = 1,2,\ldots M \qquad (1.5.25)$$

Signals $s(t)$ can be baseband, sinusoidal or signals of other carrier.

Vectorial representation of digital signals

Knowing the elements of the signal set a vector space of dimension D can be defined in which each signal corresponds to a vector. It can be shown that $D \leq M$. There is a unique relationship between signal vectors and signal waveforms:

$$s_i(t) \Leftrightarrow \mathbf{s}_i; i = 1,2...,M$$
$$\mathbf{s}_i = (s_{i,1}, s_{i,2}...s_{i,D}); s_i(t) = \sum_{j=1}^{D} s_{i,j}\varphi_j(t) \Leftrightarrow s_{i,j} = \int_0^T s_i(t)\varphi_j(t)dt \qquad (1.5.26)$$

$\varphi_j(t)$; j = 1,2,…D are the orthonormal base functions.

It can be shown that absolute square of the signal vectors is equal to the energy of the corresponding signal wave forms. It can also be shown that the band width occupied by a signal set is proportional to its dimensionality.

Vectorial treatment of signals simplifies considerably their discussion: a modulation system can be given and investigated as a geometrical structure. Vectorial treatment is in any case possible: a vector space and a signal vector set can be defined for M baseband or carrier, sinusoidal or more general finite energy signals. On the other hand note that in many practical cases carrier frequency of each signal is the same; in that case this general vectorial treatment coincides with the usual phasor representation of sinusoidal signals.

Without a more detailed discussion it is mentioned that in an AWGN (Additive White Gaussian Noisy) channel error probability depends only on the vector constellation and does not depend on the particular signal shapes. And, it depends only on the differences of signal vectors.

Binary modulation systems

Signals of digital sources are virtually always binary (M=2) and wave shapes are very simple: NRZ (Non Return to Zero) pulses. Thus a digital signal stream is composed of a sequence of “high” and “low” voltage levels. Level transition can occur

only at instances $t=kT$. It is reasonable to regard this sequence as an analog signal and apply one or an other of methods described in the previous section. This results in modulations described in the sequel.

i. Binary Amplitude Shift Keying (ASK)

Signal waveforms are:

$$s_1(t)=\sqrt{2}A\cos\omega_c t;\ s_1(t)=0;t\in(0,T) \qquad (1.5.27)$$

This is a one-dimension signal set (D =1). The single base function is:

$$\varphi(t)=\sqrt{2/T}\cos\omega_c t \qquad (1.5.28)$$

Note that this is the digital form of AM-DSB-NSC with h = 1

ii. Binary Frequency Shift Keying (FSK)

Signal waveforms and base functions are

$$\begin{aligned} &s_1(t)=\sqrt{2}A\cos\omega_c t;\ s_2(t)=\sqrt{2}A\cos(\omega_c+\delta\omega)t;\ t\in(0,T) \\ &\varphi_1(t)=\sqrt{2/T}\cos\omega_c t;\ \varphi_2(t)=\sqrt{2/T}\cos(\omega_c+2\pi/T)t;\delta\omega\geq 2\pi \end{aligned} \qquad (1.5.29)$$

This is the digital version of FM; as seen dimensionality is 2. If $\delta\omega=2\pi/T$, signal vectors are orthogonal to each other forming a binary *orthogonal signal set.*

iii. Binary Phase Shift Keying, PSK

Signal waveforms and base functions are

$$\begin{aligned} &s_1(t)=\sqrt{2}A\cos\omega_c t;s_2(t)=\sqrt{2}A\cos(\omega_c t+\Phi);t\in(0,T) \\ &\varphi_1=\sqrt{2/T}\cos\omega_c t;\varphi_2=\sqrt{2/T}\sin\omega_c t \end{aligned} \qquad (1.5.30)$$

In the general case this is also a 2-dimension signal set. However, if $\Phi=\pi$ s_1=-s_2 and D=1. In this case it can be regarded as the digital version of AM-DSB-SC.

Signal sets in which s_1=-s_2 are often called *antipodal*. It can be shown that in the binary case an antipodal signal set is of minimal error probability.

Figure. 1.5.6 shows the vector constellation of these modulation schemes.

Demodulation, decision and error probability of digital signals

Task of a system transmitting digital signals is to decide: which among the M possible signals was actually transmitted (and not to transmit signals with minimal distortion). To this *decision* operation receiver must “know” the possible signal

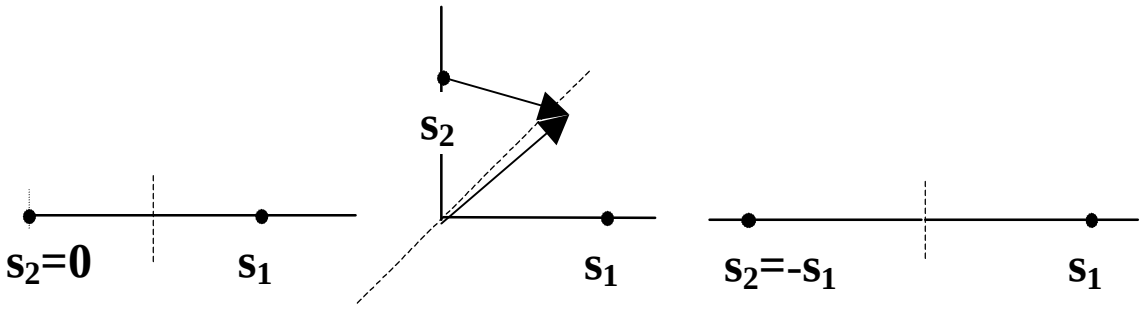


Figure 1.5.6 ASK, orthogonal FSK and antipodal PSK signal sets. Dashed lines show decision thresholds. In FSK a noise vector causing erroneous decision is also shown as well as the signal+noise vector

shapes (i.e. these must be stored in the receiver). The decision circuit determines: received signal possibly corrupted by noise, by distortion, by interference etc. is the most "likely" to which one.

More precisely: let the received signal be $r(t)$. Probability of error is minimal if we decide on the signal for which

$$\Pr\{s_i(t)|r(t)\} = \max; i = 1,2,\ldots M \qquad (1.5.31)$$

If a-priori probabilities are equal we must decide on the signal vector closest to the received one; this is assumed in Figure. 1.5.6.

Knowledge of the signal set in the receiver includes also that of the *phase* of the received signal. As phase cannot be stored in advance in the receiver, phase information must be extracted from the received signal. A demodulator recovering the phase of the received signal is called a *coherent demodulator*. Error probability of the three modulation schemes presented above if demodulated coherently is

$$P_{ePSK} = \frac{1}{2}\operatorname{erfc}\sqrt{\frac{E}{N_0}}; P_{e,FSK} = \frac{1}{2}\operatorname{erfc}\sqrt{\frac{E}{2N_0}}; P_{e,ASK} = \frac{1}{2}\operatorname{erfc}\sqrt{\frac{E}{4N_0}}; \qquad (1.5.32)$$

where E is the energy of one bit (peak energy in the case of ASK) and N_0 is the noise spectral density.

Bandwidth occupied by digital signals; multilevel modulation

In the case of radio transmission the main cost factor of the system is the occupied bandwidth. As support of the time functions is finite, the support of their Fourier transform is infinite and thus the occupied bandwidth is in theory infinite. We may regard as *practically* occupied bandwidth that band in which a given percentage

of the signal energy (90%, 95%) is contained. This band in Hz – more precisely: the band occupied by one dimension – is in the order of 1/T. E.g. the occupied band of an STM-1 signal stream of 155 Mbit/s is about 155 MHz if binary modulation is applied. It is desirable to decrease that what makes necessary the increase of the signal time. This is only possible if rather than apply bit-by-bit transmission several bits – *n* – are united into one *symbol*. Then duration of a symbol is *n.T* and, if dimensionality is not increased, the occupied bandwidth will decrease by a factor *n*. On the other hand while one bit has two possible states, the number of states of *n* bits is $M=2^n$. This needs an *M*-ary modulation scheme.

If in a signal space of 1 or 2 dimensions *M* rather than 2 vectors are to be located these will be much closer to each other; consequently a lower noise will cause error in the decision. Or, to avoid that, vector magnitudes – i.e. signal energy – must be increased. So we can state: in digital transmission the occupied bandwidth can be decreased in any proportion but then signal energy must be increased.

It can be shown that M-ary PSK is optimal if $M < 7$. If $M > 6$ modulation of amplitude *and* phase results in lower probability of error. Often MQAM (*M*-ary Quadrature Amplitude Modulation), close to optimal, is applied. As an example constellation of 16QAM is shown in Figure. 1.5.7. Probability of error of MQAM in an AWGN channel is

$$P_e = 2\left(1-1/\sqrt{M}\right)\operatorname{erfc}\frac{\sqrt{E_{peak}}}{\sqrt{2N_0}\left(\sqrt{M}-1\right)} \qquad (1.5.33)$$

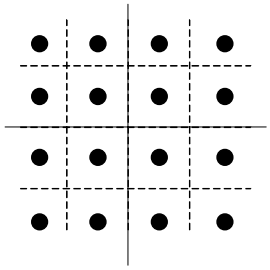


Figure 1.5.7 16QAM signal set. Dashed lines show decision thresholds

Continuous phase modulations (CPM)

These modulation schemes have three advantageous properties what lead to their wide application.

- They apply frequency band in an economic way;
- they have coding gain;
- in contrast to other frequency-economic modulation schemes (like e.g. QAM) they are not sensitive to nonlinear distortions.

In what follows we discuss briefly these properties.

It can be shown that spectral density of a random bit stream is proportional to the Fourier-transform-squared of the elementary signal. Further, the Fourier transform of a continuous function the first q-1 derivatives of which are also continuous, but the q-th is discontinuous decreases as $1/\omega^{q+1}$. (E.g. spectral density of a signal stream of constant amplitude and continuous phase decreases at least as ω^{-4}.) So if exponent q is sufficiently high it can be achieved without any filter that spurious spectral component in the band of an adjacent channel is sufficiently low. This is particularly important in narrow band transmission.

Further, continuity of the phase introduces a certain degree of redundancy. This redundancy can be applied for error correcting decoding.

Finally note that there is a nonlinear distortion if a signal on non-constant envelope is transmitted on a device of nonlinear characteristic. Consequence of nonlinear distortion can be increase of error probability and spreading of spectrum. In CPM

- information is carried by the phase;
- amplitude is constant and phase is continuous.

Thus nonlinear characteristic does not cause distortion. So neither increase of error probability nor spreading of spectrum is caused by nonlinearity.

Coded modulation schemes

In a modulated transmission system in order to decrease occupied band, it is obvious (at least now as discovered) to apply redundant states rather than applying redundant symbols. E.g. to achieve a code rate 2/3 a possibility could be: in QPSK each 2-bit data symbol is coded in a 3-bit code symbol and this transmitted via 8PSK.

This solution although not increasing occupied bandwidth would not be very advantageous: 8PSK transmission requires about 4 dB higher power than QPSK, as vectors are closer to each other. Therefore the above method would require a code of coding gain significantly higher than 4 dB; this would be needed to compensate the 4 dB loss and still yielding resulting gain.

This difficulty is overcome by partitioning of the signal set as done in Trellis Coded Modulation (TCM). In Figure. 1.5.8 partition of an 8PSK signal set into four PSK sets is shown. Partial signal sets as well as binary numbers next to the signal vectors show how the 3-bit code vectors are mapped to the points of the 8-state signal space. Regarding the figure it is obvious how much the vectors of the partial signal sets are farer from each other than those of the complete 8PSK signal space.

Functioning of an 8TCM scheme can be described as follows: first two bits of the bit-triple show the actual antipodal partial signal set and the third one shows which is the actual bit of that. Further: signal distance in this antipodal set is rather high and so correct decision of bit number 3 can be achieved with high probability. To increase the probability of correct decision of the other bits convolutional coding of sufficiently high constraint length is applied. Code rate of this is 1/2 in the 8PSK example.

The above example corresponds to the rule applied generally. I.e.:

- depending on the available bandwidth the number of bits of the uncoded symbols, n is determined;
- $k = n-1$ bits are convolutionally encoded, coding rate is $k/(k+1)$; the n-th (say the last) bit is left uncoded;
- resulting codewords of $m=n+1$ bits are mapped onto the 2-dimensional 2^{n+1} state signal space;
- to realize the last step efficiently we chose the binary partial signal sets of maximal minimum distance.

If the resulting number of levels is more than 8, QAM rather than PSK is applied.

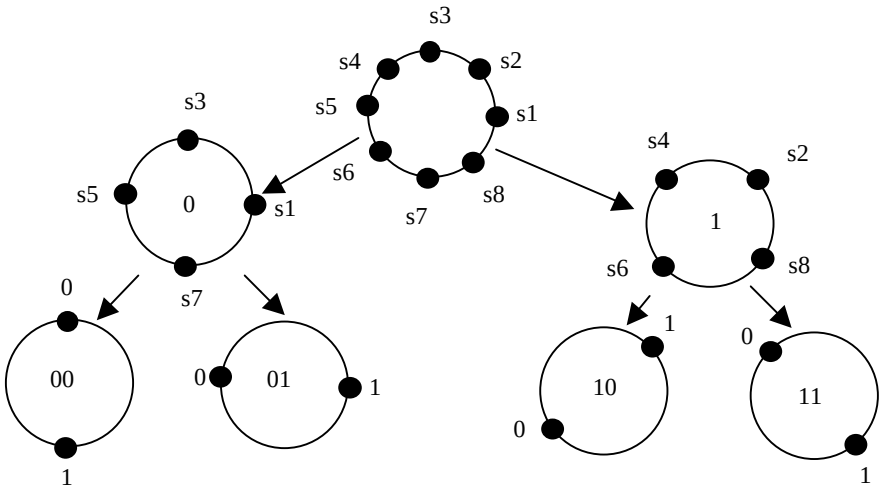


Figure 1.5.8 Partition of an 8PSK signal set to four antipodal ones

1.5.7. Modulation in the optical domain

Optical frequency band differs in many aspects from the electrical band. Bandwidth of the transmission medium is by several orders of magnitudes higher. Nearly exclusively waveguide-bound links are applied, thus users don't interfere with each other. Mainly due to these two reasons virtually always binary transmission is applied.

A further characteristic of the optical band makes plausible and virtually exclusive a special modulation method, not applied in the electrical domain: *intensity modulation*. Light intensity is a synonym of power or power density; thus it is the optical *power* what is proportional to the electrical modulating signal – to voltage or more often to current. This is justified by the fact that while electrical *current* is composed of elementary particles, i.e. electrons, elementary particles of optical *power* is composed of photons. In any optical modulator, demodulator number of photons and electrons are proportional to each other; this leads to the proportionality of electrical current and optical power.

This fact has a conceptional consequence not having, however, significant practical consequences. I.e.: as information is carried by the optical power, transmission is essentially nonlinear, namely optical signal (i.e. field strength) is proportional to the square root of the electrical signal. This ideal square root operation has really no practical consequence as optical to electrical conversion (i.e.

demodulation) is equivalent to an ideal squaring operation, from the same reason. It would have significance in optically wide band transmission; this, however, belongs even today still into the field of immature futurology.

Particularities of intensity modulation are similar to those of AM-DSB-NSC. Signal – in this case optical power – must be positive; modulated light is added to an unmodulated carrier; modulation index must be less than 1.

Although application of intensity modulation is nearly exclusive, optical PSK and FSK were also investigated, further also Polarization Shift Keying (PolSK) a scheme the electrical counterpart is not used. These are much more complicated than intensity modulation while insuring a higher capacity to the optical channel. Present-day technology does not yet require their application.

1.6. Radiowaves

Lajos Nagy dr., author

Gyula Veszely dr., reviewer

1.6.1. The radio spectrum

The electromagnetic spectrum spreads over the frequency range from near zero to 10^{23} Hz. In this very wide band take place the radio waves and also the light.

The radiospectrum is a part of the electromagnetic spectrum at which frequencies the waves can generated, radiated and received efficiently and therefore will be applicable for different radio services.

The part of the electromagnetic spectrum below 3000 GHz concerns radio waves. The radio waves propagate of radiation from an antenna without man-made ducting.

1.6.2. Wave propagation modes

The basic propagation mode for terrestrial, satellite-satellite or earth-satellite links is the free space propagation mode between transmitter and receiver antennas. This mode is typical for the microwave band.

For two terrestrial radio terminals within line-of-sight, there may be often be a ground-reflected ray in addition to direct ray. This is the two way propagation mode.

At low frequencies above ground with good conductivity propagates the surface (or ground) waves along the Earth's surface.

The ionosphere refract the radio waves from the straight line, which depend on the ionosphere behaviour and frequency. The ionospheric propagation mode forms from rays bending back to the Earth surface.

The tropospheric scatter propagation mode accomplishes by randomly distributed small-scale spatial fluctuations of refractive-index about the local mean

value which cause signal levels always to be present at large distances beyond the horizon.

1.6.3. Propagation characteristics of wavebands and system parameters

In this section, consideration is given to each decade waveband, taking into account not only the propagation aspects but also a few of the system considerations that aply.

At ELF (below 3 kHz) and VLF (3-30 kHz) bands the ionosphere forms upper boundary and Earth surface forms lower boundary to waveguide propagation. Difficult to make transmitter antennas with good efficiency.

Typical ELF services are short-range (e.g. in water, between divers) and long-range (to water) submarine communications; ultrastable worldwide communication, remote sensing under ground. The transmission rate may be only about 1 bit/s.

Typical VLF services are world-wide telegraphy with ships; fixed services over long distances; navigational aids (e.g. Omega); electrical storm warning systems and time standards.

At LF (30-300 kHz) and MF (300-3000 kHz) below 100 kHz ground waves which follow Eart curvature; sky wave becomes distinct from ground wave above this.

Typical LF services are the long-distance communication with ships; fixed services over long distances; broadcasting; radionavigational aids.

In the MF band antennas can be directional using multi-elements; L and T aerial as transmitter and ferrite-loaded coil as receiver antennas applied. Typical MF services are broadcasting; radionavigation; some land, maritime and aeronautical mobile; some fixed service.

At the HF (3-30 MHz) band ionospheric sky wave appears only beyond skip distance; surface wave are present only at short distance, more over sea.

The antennas applied in the band are log-periodic array antennas (vertical or horizontal0, vertical whip antennas or horizontal dipole arrays.

Typical services of the HF band are fixed point-to-point; land (within skip distance), maritime and aeronautical mobile; longdistance broadcasting.

The effect of the atmosphere in the VHF (30-300 MHz) band is the refraction and reflection by refractive-index irregularities producing transhorizon path; some sporadic E, ionospheric-scatter, Faraday rotation and ionospheric scintillation appear on Earth-space paths.

Terrain influences at VHF are screening by major hills, but some diffraction into valleys; surface reflections off large areas causing multipath effects on the line-of-sight paths. Typical antennas are multi-element dipole (Yagi) antennas, helixes.

Typical services of VHF band are sound and television broadcasting; land, aeronautical and marine mobile; portable and cordless telephones; aeronautical radionavigation beacons.

In the UHF (300-3000 MHz) band the atmospheric influences are refraction effects; reflection from layers at lower frequencies; ducting possible at higher frequencies; refractive-index fluctuations - forward scatter appear beyond horizon above 500 MHz.

The terrain influence is screening by hills and collections of buildings.

The typical antennas are multi-element dipole (Yagi) antennas; parabolic dishes for higher frequencies.

The typical UHF services are TV broadcasting; some aircraft navigation, landing systems; surveillance and secondary radars; mobile services; cellular radio systems.

At SHF (centimetric waves, 3-30 GHz) rain, hail snow cause very variable attenuation with frequency. High gain parabolic dishes and horns are applied. The typical services are fixed terrestrial point-to-point, point-multipoint; fixed satellite systems and radars.

At EHF (30-300 GHz, millimetric waves) and (300-3000 GHz, submillimetric waves) bands rain, hail shnow cause severe attenuation and the absorption by atmospheric gases becomes important. High gain antennas are at EHF band parabolic dishes, at submillimetric band mirror or lens antennas are applied. The

possible systems in the bands are short line-of-sight communication links and remote sensing.

1.6.4. Electromagnetic waves

The propagation modes in the radio technique can be divided into two parts based on the physical principle, these are the space and surface propagation modes.

The space waves propagate breaking away from the Earth's surface and the wave propagates toward the direction of the vectorial product of the electric and magnetic field vector, the Poynting vector. In the farfield of the sources these waves are plane waves, which can be expressed from the Maxwell's equation described for source free region.

1.6.4.1. Plane wave solution

The I and II Maxwell's equations for source free region in free space (ε_0,μ_0):

$$rot\mathbf{H} = j\omega\varepsilon_0\mathbf{E} \qquad (1.6.1)$$

$$rot\mathbf{E} = -j\omega\mu_0\mathbf{H} \qquad (1.6.2)$$

After few steps for the magnetic field strength the following homogeneous Helmholtz equation is obtained:

$$\frac{\partial^2 H_{x,y,z}}{\partial x^2} + \frac{\partial^2 H_{x,y,z}}{\partial y^2} + \frac{\partial^2 H_{x,y,z}}{\partial z^2} + \omega^2\varepsilon_0\mu_0 H_{x,y,z} = 0 \quad (1.6.3)$$

To find the magnetic field solution we substitute the (1.6.4) and (1.6.5) into the Eq. (1.6.3):

$$H_x = H_z = 0 \quad (1.6.4)$$

$$H_y = H_y^0 \cdot e^{-j\beta z} \qquad (1.6.5)$$

The Eq. (1.6.3) is satisfied if:

$$\beta^2 = \omega^2\varepsilon_0\mu_0 \quad (1.6.6)$$

To find the electric field corresponding to this plane-wave solution we substitute Eq. (1.6.5) into the I Maxwell's equation:

$$\mathbf{E} = \frac{1}{j\omega\varepsilon_0} rot\mathbf{H} = \frac{-1}{j\omega\varepsilon_0}\mathbf{e}_x \frac{\partial}{\partial z} H_y = \frac{-\mathbf{e}_x}{j\omega\varepsilon_0} H_y^0 (-j\beta) e^{-j\beta z} = E_x \mathbf{e}_x \quad (1.6.7)$$

It is evident that the electric and magnetic field vectors are perpendicular with the same phase and the ratio of their amplitudes is a constant, namely

$$\frac{E_x}{H_y} = \frac{\beta}{\omega\varepsilon_0} = \sqrt{\frac{\mu_0}{\varepsilon_0}} = 120\pi \quad (1.6.8)$$

This ratio is called the free-space wave impedance.

1.6.4.2. Polarization

Polarization of radiated wave is defined as "that property of a radiated electromagnetic wave describing the time varying direction and relative magnitude of the electric-field vector; specifically, the figure traced as a function of time by the extremity of the vector at fixed location in space, and the sense in which it is traced, as observed along the direction of propagation." Polarization then is a curve traced by the end pouint pf the vector representing the instantaneous electric filed. The field must be observed along the direction of propagation.

In general the figure that the electric field traces is an ellipse, and the field is said to be elliptically polarized. Linear and circular polarizations are special cases of elliptical, and they can be obtained when the ellipse becomes a straight line or a circle, respectively. The figure of the electric field is traced in clockwise (CW) or counterclockwise (CCW) sense. CW rotation of the electrical field vector is designated as right-hand polarized and CCW as left-hand polarized.

1.6.4.3. Reflection

The reflection take place when two different types of medium are separated by a surface which is infinite large and smooth. The amplitude, phase and polarization of the reflected wave is determined by the material parameters of the medias and the surface irregularity. When the surface is perfectly smooth the specular reflection take place and for plane incident wave the reflected wave is also plane propagating the energy in one discrete direction. This ideal case can be well described by the Snell-

Descartes law for lossless dielectrics, which can generalize using complex ε and μ for lossy dielectrics.

The reflection coefficient is investigated for two halfspace and the following two polarization. (Figure. 1.6.1.)

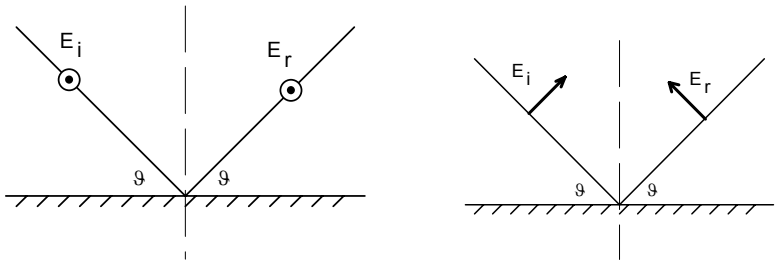


Figure 1.6.1. Horizontal polarization *Vertical*

The reflection coefficient is defined as the ratio of the amplitude of incoming and reflected electrical fieldstrength.

$$\Gamma = \frac{E_r}{E_i} \qquad (1.6.9)$$

The reflection coefficient for horizontal polarisation

$$\Gamma_h = \frac{\sin\vartheta - \sqrt{\varepsilon^* - \cos^2\vartheta}}{\sin\vartheta + \sqrt{\varepsilon^* - \cos^2\vartheta}} \qquad (1.6.10)$$

The reflection coefficient for vertical polarisation

$$\Gamma_v = \frac{\varepsilon^* \sin\vartheta - \sqrt{\varepsilon^* - \cos^2\vartheta}}{\varepsilon^* \sin\vartheta + \sqrt{\varepsilon^* - \cos^2\vartheta}} \quad (1.6.11)$$

where

$\varepsilon^* = \varepsilon' + \dfrac{\sigma}{j\omega\varepsilon_o} = \varepsilon' - j\,60\lambda\,\sigma = \varepsilon' - j\,\varepsilon''$ the complex dielectric constant.

In the radiowave propagation modelling the ground reflection is the main task. Let us introduce the ground reflection coefficients for two frequencies.

At ϑ_B incoming angle the $|\Gamma|$ for vertical polarisation has minima.

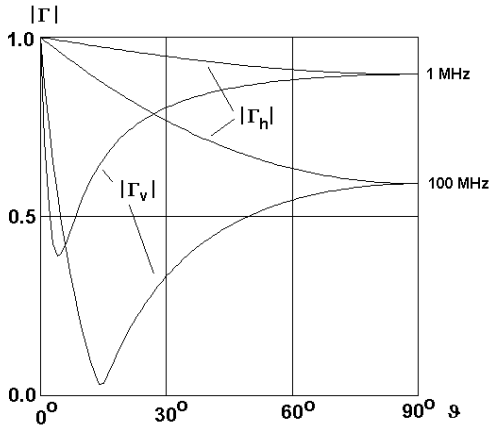


Figure 1.6.2. Magnitude of Earth reflection coefficient

If σ=0, then $tg\vartheta_B = 1/\sqrt{\varepsilon'}$. At this angle $\Gamma_v = 0$ and ϑ_B is the Brewster angle, if $\sigma \neq 0$, then $tg\vartheta_B \cong 1/\sqrt{\varepsilon'}$ and ϑ_B is the a pseudo Brewster angle.

1.6.4.4. Diffraction

The free space propagation is often obstructed by Earthly obstacles. When a wavefront encounters an obstacle or discontinuity that is comparable or not large compared to the wavelength, then the Huygens principle is often useful in giving an insight into the problem and in providing a solution. The principle suggests that each point on a wavefront acts as the source of a secondary wavelet and that these wavelets combine to produce a new wavefront in the direction of propagation. The effect of such obstacles is described using the diffraction theory.

The field strength at the point R in Figure. 1.6.3 is determined as the sum of all the secondary Huygens sources in the plane above the obstruction and can be expressed by the following Fresnel integral as

$$\frac{E}{E_o} = \frac{1}{1-j}\int_{\nu_o}^{\infty} \exp(-j\frac{\pi}{2}\nu^2)d\nu \quad (1.6.12)$$

where the radius of the first Fresnel ellipse r_1 and the knife edge relative height ν_o:

$$r_1 = \sqrt{\frac{\lambda d_1 d_2}{d_1 + d_2}} \quad \rightarrow \quad \nu_o = \sqrt{2}\frac{h}{r_1} = h\sqrt{\frac{2(d_1 + d_2)}{\lambda d_1 d_2}} \qquad (1.6.13)$$

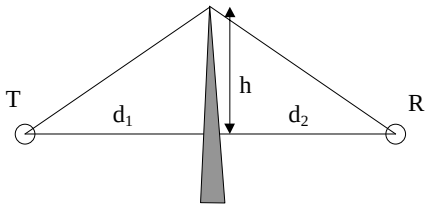
Figure. 1.6.4. shows the diffraction loss in dB relative to the free space loss.

1.6.4.5. Scattering

Scattering can be described as disorganised reflection from a rough surface. The Rayleigh roughness criterion provides a useful guide as to whether a surface should be considered smooth or rough as far as reflection is concerned. If the phase difference between the rays reflected from different points of the surface is smaller than $\pi/2$, the surface will support specular reflection and should be considered smooth. If the difference is much greater, the surface is rough and will produce scattering. The path difference from the phase criterion is $\lambda/4$.

The path difference caused by the ground roughness is $\Delta l = 2 \cdot \Delta h \cdot \sin \vartheta^i$ from the Figure. 1.6.5. and therefore the maximal roughness allowed by the Rayleigh criterion is $\Delta h = \frac{\lambda}{8 \sin \vartheta^i}$.

By rough surfaces the height distribution of the surface is modelled by Gaussian and the scattering loss ρ_s is given by



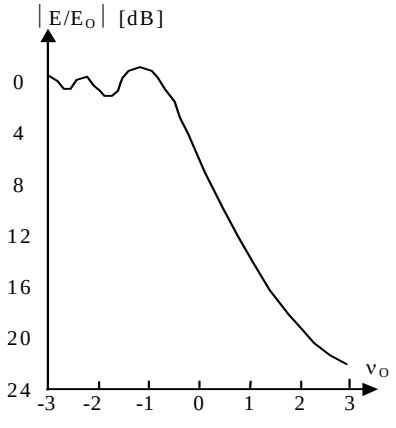


Figure 1.6.3. The geometry of knife-edge diffraction Figure 1.6.4. Diffraction loss over a single knife-edge as a function of the obstacle relative height

$$\rho_s = \exp\left[-8 \left(\frac{\pi \sigma_s \sin \vartheta^i}{\lambda} \right)^2 \right] \quad (1.6.14)$$

where σ_s is the surface height distribution standard deviation.

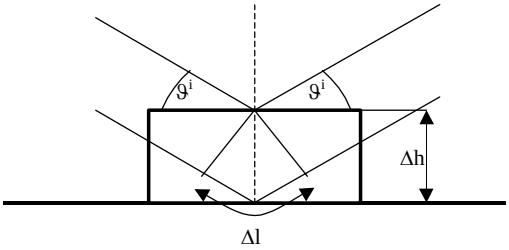


Figure 1.6.5. Geometry for Rayleigh roughness criterion

The reflection coefficient of the rough surface is

$$\Gamma_{rough} = \rho_s \cdot \Gamma_{plane} \qquad (1.6.15)$$

1.6.5. The effect of the Earthly atmosphere

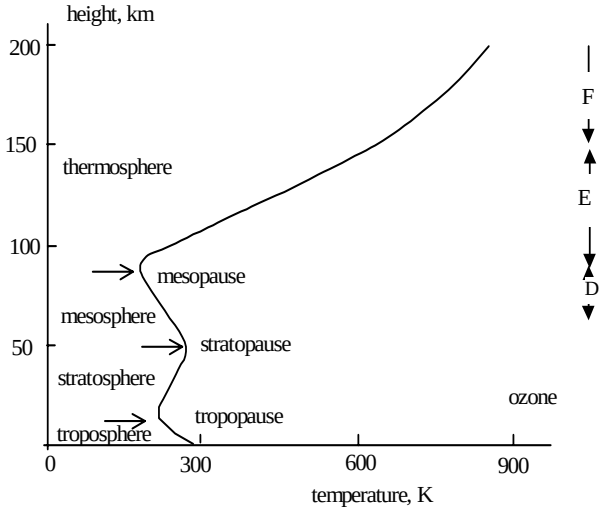


Figure 1.6.6. Regions of the Earth's atmosphere

The power density from an antenna decreases during propagation because of two different phenomena:

1. the wave diverges from the antenna;
2. the wave is progressively absorbed or scattered by the medium.

The second phenomenon is caused mainly by the two following effect.

1. molecular absorption by atmospheric gases;
2. absorption and scatter by liquid and solid particles in the atmosphere particularly by raindrops.

These effects begin to appear at frequencies above few GHz, and increase rapidly with increasing frequency. Moreover, particles can lead to a modification in the polarization of the incident wave.

1.6.5.1. Attenuation by Atmospheric Gases

Since nitrogen has no absorption bands in the radio frequency range, molecular absorption is due almost entirely to oxygen and the water vapour.

At frequencies less than 350 GHz *oxygen* has an isolated absorption line at 118,74 GHz and a series of very close lines between about 50 GHz and 70 GHz. In the lower part of the atmosphere these lines behave continuous absorption band.

At frequencies lower than 350 GHz, *water vapour* shows three absorption lines, at frequencies of 22.3 GHz, 183.3 GHz and 323,8 GHz. Furthermore, at submillimetre wave range and infrared frequencies there are many lines, some of them very intense.

For very low concentrations of water vapour attenuation can be treated as being proportional to concentration.

Figure 1.6.7. shows the attenuation values for atmospheris gases. For water vapour, we chose a value equal to 7.5 g/m^3, which corresponds to 1% water vapour molecules mixed with 99% molecules of dry air. This value corresponds, at ground level, to 50% relative humidity if the temperature is 16.5^{o}C or 75% at 10^{o}C.

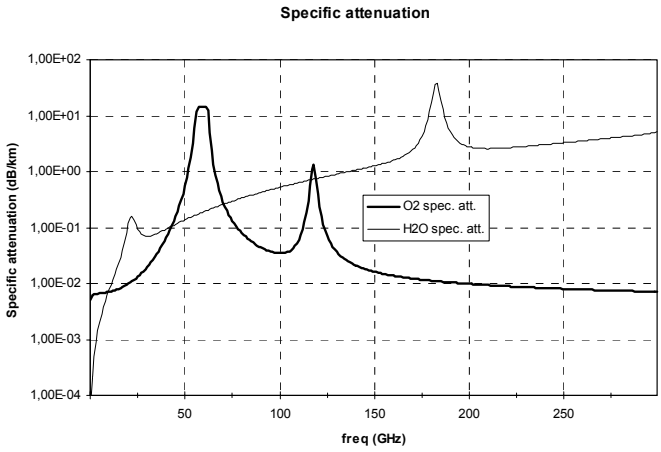


Figure 1.6.7. Attenuation of the atmospheric gases

1.6.5.2. Attenuation by moisture

Usually the attenuation caused by the rain is the main effect investigated. Usually only the rain rate R in millimeters of water per hour can be easily measured. A light drizzle corresponds to a rain rate of 0.25 mm/h, light rain corresponds to R=1mm/h, moderate rain to 4 mm/h, heavy rain to 16 mm/h, and cloud bursts up to many centimeters per hour. The drop-size distribution is a function of rain rate, with a greater concentration of large drops occuring for heavy rain. Marshal and Palmer proposed the following empirical distribution formula:

$$N(a) = N_0 e^{-\Lambda a} \qquad (1.6.16)$$

where $N_0 = 1.6 \times 10^4 \ mm^{-1}/m^3$ and $\Lambda = 8.2 R^{-0.21} mm^{-1}$,

a is the radius of the raindrops in mm.

From the point of view of the communications engineer what is needed is a relatively simple formula relating specific attenuation to rain rate, frequency, and temperature. Such formula can be described in the following form:

$$A = cR^b \quad [dB/km] \quad (1.6.17)$$

where c and b are constants that depend on frequency and temperature of the rain.

Some representative curves of specific attenuation of rain at frequencies of 1, 3.5 and 10 GHz as a function of rain rate were computed using (1.6.17) and are shown in Figure.1.6.8.

Another effect of rain that can be significant in dual polarized systems is the *depolarization of radio waves*. That is, some of the energy in each polarization is transformed into energy of the opposite (orthogonal) polarization.

Viewed from the radar theory point of view, raindrops also cause bistatic scattering. It could be of importance in the case of space division multiplexing since it can cause cochannel interference to occur when the signal in one beam gets scattered and is received by a receiver from another sector.

The attenuation of microwaves and millimeter waves by *fog* is governed by the same fundamental equations as attenuation by rain. The main difference is that fog is suspended mist of very small water droplets with radii in the range of 0.01 to 0.05

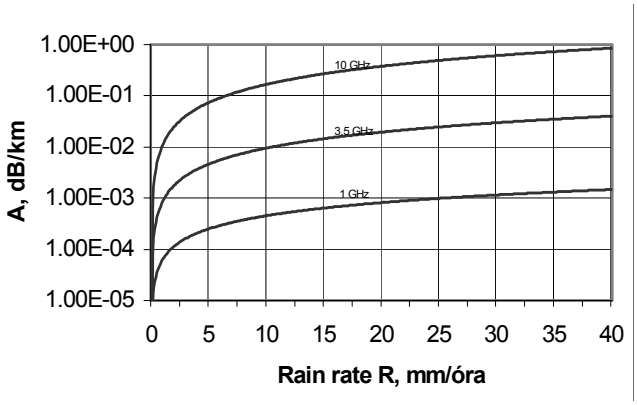


Figure 1.6.8. Attenuation by rain at 1, 3.5 and 10 GHz as a function of rain rate

mm. For frequencies below 300 GHz the attenuation by fog is essentially linearly proportional to the total water content per unit volume at any given frequency. The upper level for water content is around 1 g/m^3, with the content usually considerably less than this for most fogs. A concentration of 0.032 g/m^3 corresponds to a fog that is characterized by an optical visibility of around 600 m. A concentration of 0.32 g/m^3 corresponds to an optical visibility range of around 120 m. The attenuation by fog in dB-s per kilometer as a function of frequency is shown in Figure. 1.6.9. for the two concentration levels mentioned above.

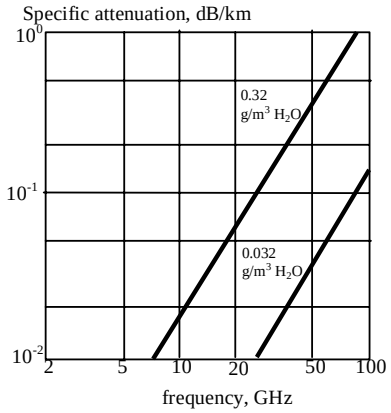


Figure 1.6.9. Attenuation by fog as a function of frequency for two different concentrations

At a frequency of 300 GHz the attenuation in the more dense fog is still only about 1 dB/km. Hence, for communication link designs with sufficient signal margin built in to overcome the attenuation by rain, the attenuation by fog will not be a limiting factor.

Snow and hail consist of a mixture of ice crystals and water in many instances, so the attenuation is strongly dependent on the meteorological conditions. Furthermore the shape of snow and ice crystals is so varied that the calculation of absorption by a single typical particle is a formidable task, it indeed a typical particle can even be defined.

Attenuation of microwaves in dry snow is at least an order of magnitude less than in rain for the same precipitation rate. However, attenuation by wet snow is comparable to that in rain and may even exceed that of rain at millimeter wavelengths. Even in dry snow, measurements have shown that the attenuation of 0.96 mm radiation may be greater than in rain with the same precipitation rate. Measurements have shown an attenuation of around 2 dB/km at 35 GHz for wet snow and a precipitation rate of 5 mm/h. For dry snow the attenuation is two orders of magnitude less.

1.6.5.3. Atmospheric refraction

If we ignore special cases such as links between satellites, all radio path travel through the atmosphere. The lower part of the atmosphere, the troposphere has the main role in the refraction of the radio waves propagating through. Here the atmospheric gases have the concetration high enough to refract the radio waves from the straight line.

This effect is caused by the spatial variation of the troposphere refractivity.

The radio refractive index of the trophosphere is due to the molecular constituents of the air, principally nitrogen, oxygen, carbon dioxide and water vapour. The value of refractivity deviates from unity because of:

- The polarisability of these molecules due to the incident electromagnetic field
- molecular resonances

The latter effect is limited to narrow frequency bands (for example around 22GHz and 60 GHz).

The first effect is independent of frequency at the frequencies of interest (up to millimetre waves).

A typical value of the air refractivity at the Earth's surface $n \cong 1.0003$.

Because of the closeness of n to unity, it is usual to work with the refractive index *N* defined by:

$n=1+10^{-6}N$ (1.6.18)

When dealing with radio waves the refractive index of air is given by the following approximate formula, which is adopted by the ITU-R.

$$N=(1-n)10^{6}=77.6\frac{p}{T}+3.73\cdot 10^{5}\frac{e}{T^{2}} \qquad (1.6.19)$$

p is the total pressure in millibars;

T is the absolute temperature in Kelvins;

e is the partial pressure of the water vapour in millibars.

These meteorological parameters vary with altitude.

In reference atmosphere for refraction the n decreases with altitude and therefore the radio waves refracted toward the Earth. (Figure. 1.6.10)

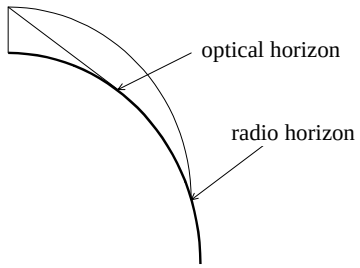


Figure 1.6.10. Refraction of radio waves

At sea level the value of *N* is around 300.

The path of the ray propagating through the atmosphere can be expressed using the Fresnel's law. If we introduce the Earth's radius coefficient *K* the Straight-line ray propagation relative to the effective Earth radius can be arranged by setting the Earth radius to $K \cdot R_0$ for radiowave propagation tasks.

1.6.6. Free space radio link

Radio propagation is a subject where deterministic analysis can only be applied in a few, rather simple cases. The extent to which these cases represent

practical conditions is a matter for individual interpretation, but they do give an insight into the basic propagation mechanisms and establish bounds.

With using the plane wave propagation mechanism from the 1.6.4.1. section in the far field of the antennas and for point sources the attenuation of the free space propagation path is determined.

If a transmitting antenna is located in free space, i.e. remote from earth or any obstructions, then if it has a gain G_T in the direction to a receiving antenna, P_T transmitted power, the power density at a distance d in the chosen direction is

$$S = \frac{P_T G_T}{4\pi d^2} \qquad (1.6.20)$$

The available power at the receiving antenna, which has an effective area A_e is therefore

$$P_R = \frac{P_T G_T}{4\pi d^2} A_e = \frac{P_T G_T}{4\pi d^2} \cdot \frac{\lambda^2 G_R}{4\pi} \quad (1.6.21)$$

where G_R is the gain of the receiving antenna, λ is the wavelength.

For the attenuation of the radio channel (free space loss) we obtain

$$L_F = \frac{P_R}{P_T} = G_T G_R \left[\frac{\lambda}{4\pi d}\right]^2 \qquad (1.6.22)$$

The free space loss or path loss is conveniently expressed in dB and from Eq. (1.6.22) with the frequency f we can write

$$L_F = \left(G_T^{dB} + G_R^{dB}\right) - 20\log f - 20\log d + 147.6 \qquad (1.6.23)$$

It is often useful to compare path loss with the basic path loss L_B between isotropic antennas, which is

$$L_B = -32.44 - 20\log f^{MHz} - 20\log d^{km} \qquad (1.6.24)$$

1.6.7. Surface wave propagation mode

The surface wave take place along the separating surface of the well conducted Earth and air. The antennas are small compared to the wavelength and therefore the direct and reflected waves cancel each other. The electric field lines can be seen on the Figure. 1.6.11.

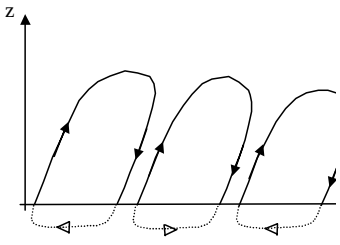


Figure 1.6.11. Propagation of surface waves

In the pure surface wave the vertical and radial components of the electric field are present, and this means that the propagation wavefront is tilted.

Surface waves can radiated efficiently using vertical monopole antenna with Earth connected ground plane to the.

Sommerfeld published in 1909 the loss factor of the surface waves.

$$E = E_0 \cdot A(p) \quad (1.6.25)$$

where

A(p) loss factor for vertical polarized surface waves

E_0 free space electrical field strength

The electrical field strength of the surface waves decrease with respect to the square of the distance from the transmitter antenna. This model for plane Earth can be applied nearer then $d[km] = 80/\sqrt[3]{f[MHz]}$, fahrer Sommerfeld has given a correction for spherical ground.

1.6.8. Tropospheric scatter

The refractive-index of the atmosphere vary in long time average regular way and can be well decribed. Randomly distributed small-scale spatial fluctuations of

refractive-index about the local mean value cause weak signal levels always to be present at large distances beyond the horizon.

Troposperic-scatter-propagation links operate in the frequency range of 200 MHz up to 10 GHz. Operation at lower frequencies is not attractive because of the cost of building antennas with sufficient gain. At higher frequencies the transmission loss becomes large. The typical distance involved in a tropospheric scatter link is a few hundred kilometers, usually not more than 1000, and the scattering common volume forms in the atmosphere usually below 10 km.

1.6.9. Ionospheric propagation

The ionosphere is the region of the atmosphere in which ionisation of gases is particularly intense, and it extends from heights of approximately 60 km to 600 km. The ionisation is caused mainly by ultraviolet radiation from the sun and slightly from the ionisation effect of the meteorits. The behaviour of the ionosphere is very closely correlated with the sun activity and can be therefore predicted using the sunspot numbers.

The ionosphere may be regarded as a low-conductivity dielectric with refractive-index always less than unity, decreasing as electron density increases and/or frequency decreases. If the frequency is increased, the height of reflection increases, and at the *critical frequency*, the ray will penetrate the ionospheric layer.

The ionospheric links can be estrablished by one or more ionospheric skips operating few tausand kilometer link distances.

References

[1.6.1] M.P.M. Hall, L.W.Barclay, M.T.Hewitt: Propagation of Radiowaves, The IEE, London, 1996.

[1.6.2] R.E. Collin: Antennas and Radiowave Propagation, McGraw-Hill Book Company, New York, 1985.

[1.6.3] L. Boithias: Radio Wave Propagation, North Oxford Academic, Publ. Ltd., London, 1987.

1.7. Teletraffic Theory

Sándor Molnár dr., author

László Jereb dr., reviewer

1.7.1. Introduction

Teletraffic theory [1.7.1], [1.7.7] is the basis for performance evolution and dimensioning of telecommunication networks. It was founded by Agner Krarup Erlang (1878-1929) [1.7.6], a Danish mathematician, at the beginning of the 20th century. The theory has been developed along with the developments of telephone networks [1.7.7] and it is an essential component in the design of traditional telecommunication networks.

Teletraffic theory has been developed together with the enormous developments of switching and networking technology in the last decades. It has been incorporating the recent advances of operation research and queueing theory. Integrating the results of different fields a continuous evolution of teletraffic theory can be observed.

Teletraffic theory deals with the application of mathematical modeling of the traffic demand, network capacity and realized performance relationships. The traffic demand is statistical in nature resulting in appropriate models derived from the theory of stochastic processes.

In this chapter first we present an introduction about the characteristics of network traffic. The nature of traffic had a strong impact on the developed teletraffic theory we have today. After we overview the basics of teletraffic theory including the notations, classification of systems and the fundamental teletraffic equations. The applications of basic teletraffic results outlined in this chapter can be found in Chapter 3.3 where teletraffic models and the teletraffic dimensioning methods are described.

1.7.2. The characteristics of network traffic

The nature of traffic in today's data networks (e.g. Internet) is completely different from classical telephone traffic and the characterization is not as simple as it was in the case of conventional POTS traffic [1.7.8], [1.7.9]. The main difference can be explained by the fact that in traditional telephony the traffic is highly *static* in nature. It was possible to find a typical user and behavior where averages simply describe the system performance adequately due to the *limited variability* of traffic characteristics.

The static nature of telephone traffic resulted in "universal laws" governing telephone networks like the *Poisson nature* of call arrivals [1.7.8], [1.7.9]. This law states that call arrivals are mutually independent and exponentially distributed with the same parameter. The Poisson call arrival model had a general popularity in the last fifty years. The great success of the Poissonian model is due to the parsimonious modeling, which is a highly desirable property in practice.

A similar "universal law" of the POTS traffic is that the call holding times follow more or less an *exponential distribution*. This model was also preferred due to its simplicity and analytical tractability in spite of the fact that the actual telephone call duration distribution sometimes deviates significantly from the exponential distribution. However, these deviations did not yield to major errors in network design thanks to the nice nature of Poisson arrival process. This is because several performance measures do not depend on the distribution but only of the average of holding time.

A dramatic change happened concerning the validity of these laws when telephone networks were used not only for voice conversations but also for FAX transmissions and Internet access. The statistical characteristics of these services are significantly different from voice calls. Especially, the call durations become much longer and more variable compared to classical voice calls. As the popularity of the Internet increased due to the success of Web, more and more people started to use the classical telephone networks for Internet access. These changes call for reviewing the old laws and present a challenge for today's teletraffic researchers.

The picture is completely different in case of data networks. All the expectations by finding similar universal laws for data traffic failed [1.7.8]. It is

because data traffic is much more variable than voice traffic. Roughly speaking, it is impossible to find a general model because the individual connections of data communication can change from extremely short to extremely long and the data rate can also be in a huge range. There is no static and homogenous nature of data traffic as it was found in case of the voice traffic. This extremely bursty nature of data traffic is mainly caused by the fact that this traffic is generated by machine-to-machine communication in contrast to the human-to-human communication.

This high variability of data traffic in both *time* (traffic dependencies do not decay exponentially fast as it was the case in voice traffic but long-term dependencies are present, e.g. in the autocorrelation of the traffic) and in *space* (distributions of traffic related quantities do not have exponential tails as it was the case in the case of voice traffic but heavy tails are very common, e.g. in distributions of web item sizes) call for new models and techniques to be developed. Statistically, the long-term dependencies can be captured by *long-range dependence* (LRD), i.e., autocorrelations that exhibit power-law decay. The extreme spatial variability can be described by *heavy-tailed distributions* with infinite variance, which is typically expressed by the Pareto distributions. The power-law behavior in both time and space of some statistical descriptors often cause the corresponding traffic process to exhibit *fractal* characteristics [1.7.8].

The fractal properties often manifest themselves in *self-similarity*. It means that several statistical characteristics of the traffic are the same over a range of time scales. Self-similar traffic models seem to be successful parsimonious models to capture this complex fractal nature of network traffic in the previous decade. However, recent research indicates that the actual data traffic has a more refined burstiness structure, which is better captured by *multifractality* rather than only self-similarity, which is a special case of *monofractality*. Multifractal traffic models have also been developed [1.7.8].

Besides the very variable characteristics of data traffic there are other factors that make predictions about data traffic characteristics more unreliable. The Internet traffic is doubling each year. This extreme traffic increase with the possible so-called “killer applications” could disrupt any predictions. However, from the history of the Internet we can identify only three “killer applications” that dramatically changed the traffic mix of the Internet (the e-mail, the web and the recently emerging Napster-like

applications) but nobody knows when we can face a popular application which will take the major role of the Internet traffic characteristics. The picture is even more complicated if we think of Quality of Service (QoS) requirements of data services which can be very different from one application to the other. Different QoS requirents generate different traffic characteristics.

To describe these different traffic characteristics in case of both stream and elastic traffic flows a number of traffic models and traffic characterization techniques have been developed. Based on a successful traffic modeling one can also hope to find successful traffic dimensioning methods for resource allocation. The most important traffic models and dimensioning methods are described in Chapter 3.3.

1.7.3. Basic concepts of teletraffic theory

In this subsection the most important teletraffic concepts are overviewed [1.7.1].

1.7.3.1. Basic notions

A demand for a connection in a network is defined as a *call,* which is activated by a *customer*. The call duration is defined as *holding time* or *service time*. The *traffic load* is the total holding time per unit time. The unit of traffic load is called *erlang* (erl) after the father of teletraffic theory.

The traffic load has the following important properties:

1. The traffic load (offered traffic) a is given by $a=ch$ (erl) where c is the number of calls originating per unit time and h is the mean holding time.
2. The traffic load (offered traffic) is equal to the number of calls originating in the mean holding time.
3. The traffic load (carried traffic) carried by a single trunk is equivalent to the probability (fraction of time) that the trunk is used (busy).
4. The traffic load (carried traffic) carried by a group of trunks is equivalent to the mean (expected) number of busy trunks in the group.

1.7.3.2. Classification of teletraffic systems

The *switching system* is defined as a system connecting between inlets and outlets. A system is called a *full availability system* if any inlet can be connected to

any idle outlet. *Congestion* is a state of the system when a connection cannot be made because of busy outlets or internal paths. The system is called a *waiting* or *delay system* if an incoming call can wait for a connection in case of congestion. If no waiting is possible in congestion state the call is blocked and the system is called as *loss system* or *non-delay system*.

A full availability system can be described by the following [1.7.1]:

1. Input process: This describes the way of call arrival process.

2. Service mechanism: This describes the number of outlets, service time distributions, etc.

3. Queue discipline: This specifies ways of call handling during congestion. In delay systems the most typical queueing disciplines are the first-in first-out (FIFO), last-in first-out (LIFO), priority systems, processor sharing, etc.

The *Kendall notation* is used [1.7.1], [1.7.3], [1.7.4] for classification of full availability systems named after David A. Kendall, a British statistician:

A/B/C/D/E-F

where *A* represents the interarrival time distribution, *B* service time distribution, *C* number of parallel servers, *D* system capacity, *E* finite customer population, and *F* is the queueing discipline. The following notations are used:

M: Exponential (Markov)

E_k: Phase k Erlangian

H_n: Order n hyper-exponential

D: Deterministic

G: General

GI: General independent

MMPP: Markov modulated Poisson process

MAP: Markov arrival process

As an example *M/M/1/*∞*//*∞-FCFS represents a queueing system with Poisson arrivals and exponentially distributed service times. The system has only one server, an infinite waiting queue. The customer population is infinite and the customers are served on a first come first served basis.

1.7.3.3. Fundamental relations

PASTA

For a Poisson arrival process (exponential interarrival times) in steady state the distribution of existing calls at an arbitrary instant is equal to the distribution of calls just prior to call arrival epochs. This relationship is called *PASTA* (Poisson arrivals see time averages) [1.7.1] because this probability is equal to the average time fraction of calls existing when observed over a sufficiently long period.

Markov property

If the interarrival time is exponentially distributed, the residual time seen at an arbitrary time instant is also exponential with the same parameter. A model with interarrival time and service time both exponentially distributed is called *Markovian model* [1.7.1], otherwise it is called *non-Markovian model*.

Little Formula

The formula $N=\lambda W$ is called the *Little formula* [1.7.1], [1.7.3], [1.7.4] where N is the mean number of customers in the system, λ is the mean arrival rate and W is the mean waiting time in the system. Note that the Little formula applies to any stationary system where customers are not created or lost in the system.

Loss Formula

The probability of an arbitrary customer being lost [1.7.5] is

$$P_{loss} = 1 - \frac{1-\phi}{\rho},$$

where ρ is the offered load and ϕ is the probability that the server is idle. The formula is called the *loss formula* and is also valid for multiserver systems with ρ being interpreted to be the mean load per server and ϕ is the probability that arbitrarily chosen server is idle.

Unfinished work vs. number in the system

In a constant service time single server system we have the following relationship between the unfinished work V_t in the system and the number of costumers X_t in the system: $X_t = \lceil V_t \rceil$. Based on this we have the following identity for the complementary distributions [1.7.5]:

$P(X_t > n) = P(V_t > n)$, for *n* an integer.

Packet loss probability vs. queue length tail probability

Consider a discrete time *G/D/1* system with fixed length packet arrivals (cells). An upper bound for the cell loss probability can be given [1.7.5] by

$$\rho P_{loss} \le P(X_t^{\infty} > K),$$

where ρ is the load, X_t^{∞} is the queue length in a hypothetical infinite capacity queue.

The generalized Beneŝ formula

Consider a service system with unlimited buffer. Assume that the system is stationary so that 0 represents an arbitrary time instant. The server capacity is 1 unit work per unit of time. The complementary distribution of the amount of work in the system at time 0 can be computed [1.7.5] by

$$P(V_0 > x) = \int_{u>0} P(\xi(u) \ge x > \xi(u+du) \qquad and \qquad V_{-u} = 0),$$

where $\xi(t)$ is defined by $\xi(t) = A(t) - t$, $t \ge 0$, and *A(t)* is the amount of work arriving to the system in the interval [-*t*,0). The result covers all realizable queueing systems and found to be very useful in teletraffic theory.

1.7.4. The M/G/1 queue

The queueing system with Poisson arrivals, general service time distributions and a single server (*M/G/1*) is a very important category in teletraffic theory. In this subsection we overview the major results related to this queueing system [1.7.1], [1.7.2], [1.7.3], [1.7.4].

The following notations are used:

W: waiting time in the queue

T: response time in the system

N_q: number of customers in the queue

N: number of customers in the system

S: service time

The average waiting time and the number of customers in the queue for the *M/G/1* queueing system are given by the following equations [1.7.2]:

$$E(W) = \frac{\rho E(S)(1+c_S^2)}{2(1-\rho)},\ E(N_q) = \frac{\rho^2(1+c_S^2)}{2(1-\rho)}$$

where ρ is the load of the queue and c_S^2 is the squared coefficient of variation of the service time, i.e. Var(S)/E^2(S).

The distribution of the number of customers in the system can be computed from the Pollaczek-Khinchin transform equation [1.7.2]:

$$G_N(z) = L_S(\lambda(1-z))\frac{(1-\rho)(1-z)}{L_S(\lambda(1-z))-z},$$

where $G_N(z) = E[z^N]$ the probability generating function for *N*, $L_X(s) = E[e^{-sX}]$ the Laplace transform for *X* and λ is the Poisson arrival rate. Based on this key equation the obtained queue length distributions for the most frequently used *M/G/1* systems are summarized below.

Queue	Queue length distribution P(*N*=*n*)
M/M/1	$(1-\rho)\rho^n$
M/H₂/1	$q(1-\alpha_1)\alpha_1^{\ n} + (1-q)(1-\alpha_2)\alpha_2^n$
M/D/1	$(1-\rho)\sum_{k=0}^{n} e^{k\rho}(-1)^{n-k}\frac{(k\rho+n-k)(k\rho)^{n-k-1}}{(n-k)!}$
M/E$_k$/1	$(1-\rho)\sum_{j=0}^{n}(-1)^{n-j}\frac{\alpha^{n-j-1}}{(1-\alpha)^{kj}}\left[\binom{kj}{n-j}\alpha+\binom{kj}{n-j-1}\right]$

Here H_2 means the *hyperexponential* distribution given by parameters α_1, α_2 and *q*. E_k refers to the *k-Erlangian* distribution given by parameters α and *k*. These two distributions are important because using them we can approximate *M/G/1* systems where the squared coefficient of variation of the service time less than or equal to 1 (*M/E$_k$/1* queues) and greater than or equal to 1 (*M/H₂/1* queues).

1.7.5. General queueing systems

General queueing systems (*G/G/n* queues) are usually difficult to solve but there are subclasses that can be handled more easily than others. For example, the *G/M/1* queueing system is less useful than the *M/G/1* in data networks but the analysis is simpler than its dual pair. A remarkable result of the *G/G/1* systems is the *Lindley's integral equation* [1.7.1], [1.7.3] which gives the stationary waiting time distribution:

$$F_W(t) = \int_{-\infty}^{t} F_W(t-v)dF_U(v),$$

where the random variable $U = S - A$ with A denoting the time between the arrivals of two consecutive customers.

1.7.6. Teletraffic techniques

Beyond the classical queueing methods there are numerous approximations, bounds, techniques to handle teletraffic systems. In this subsection we overview the most significant methods.

The *fluid flow approximation* [1.7.3] is a useful technique when in the time scale under investigation we have lots of traffic units (packets). In this case we can treat it as a continuous flow like fluid entering a piping system. We can define *A(t)* and *D(t)* to be the random variables describing the number of arrivals and departures respectively in (0,*t*). The number of customers in the system at time *t* is *N(t)=A(t)-D(t)*, assuming that initially the system is empty. By the weak law of large numbers, when *A(t)* gets large it gets close to its mean and this is the same for *D(t)*. The fluid flow approximation simply replaces *A(t)* and *D(t)* by the their means, which are continuous deterministic processes. Fluid flow models are frequently used in teletraffic systems modeling.

The fluid flow approximation uses mean values and the variability in the arrival and departure processes is not taken into account. The *diffusion approximation* [1.7.3] extends this model by modeling this variability (motivated by the central limit theorem) by normal distribution around the mean. Diffusion approximations are also

applied to solve difficult queueing systems. For example, the in the complex *G/G/1* system the queue length distribution can be obtained by diffusion methods.

An approach based on the information theory called the *maximum entropy method* [1.7.3] is often useful in solving teletraffic systems. The basis is Bernoulli's principle of insufficient reasons which states that all events over a sample space should have the same probability unless there is evidence to the contrary. The entropy of a random variable is minimum (zero) when its value is certain. The entropy is maximum when its value is uniformly distributed because the outcome of an event has maximum uncertainty. The idea is that the entropy be maximized subject to any additional evidence. The method is successfully used for example in queueing theory.

A number of other methods have also been developed like queueing networks with several solving techniques, fixed point methods, decomposition techniques, etc. Interested readers should refer to the reference list of this chapter.

References

[1.7.1] H. Akimaru. K. Kawashima: Teletraffic, Theory and Applications, Springer-Verlag, 1999.

[1.7.2] R. Nelson: Probability, Stochastic Processes, and queueing Theory, Springer-Verlag, 1995.

[1.7.3] P. G. Harrison, N. M. Patel: Performance Modelling of Communication Networks and Computer Architectures, Addison-Wesley, 1993.

[1.7.4] R. Jain: The Art of Computer Systems Performance Anaysis, Wiley, 1991.

[1.7.5] J. Roberts, U. Mocci, J. Virtamo (eds.), Broadband Network teletraffic, Springer-Verlag, 1996.

[1.7.6] E. Brockmeyer, F. L. Halstrom, A. Jensen: The Life and Works of A. K. Erlang, Acta Polytechnica Scandinavica, 1960.

[1.7.7] R. Syski, Introduction to Congestion Theory in Telephone Systems, Oliver and Boyd Ltd. 1960.

[1.7.8] W. Willinger, V. Paxson: Where Mathematics Meets the Internet, Notices of the American Mathematical Society, vol.45, no.8, Aug. 1998, pp. 961-970.

[1.7.9] J. Roberts, Traffic Theory and the Internet, IEEE Communications Magazine, January 2000.

1.8. Data protection and publicity

Iván Székely dr., author

István Vajda dr., reviewer

1.8.1. Basic models

The basis for classifying the whole range of data and information, from raw information to recorded, structured and retrievable data, and from data to contextualized and re-interpreted information, is its personal or public nature. These two basic categories cover the whole range of data transmitted and processed in telecommunication systems (Figure 1.8.1). Although this categorization differs from the logic usually applied in telecommunication and information technologies, it implies basic rules which determine the controlling of data and its technological realization. These rules are not only of legal or sociological nature: they derive from the philosophy of data controlling through its formulated basic principles and the means of legislation, regulation and self-regulation, and extend up to information and communication technologies. In a general sense they determine the basic framework of openness and secrecy in the modern data processing environment.

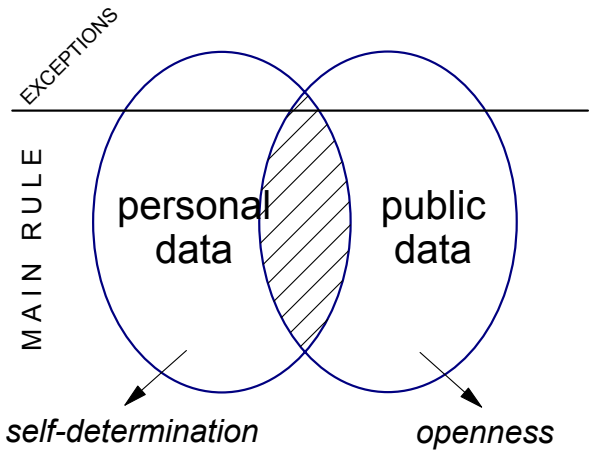


Figure 1.8.1

In Székely's basic model (reflecting a philosophical and legal-theoretical approach), as illustrated here, both basic data categories are governed by a main rule: for personal data this is self-determination and for the public data, openness. (The equivalent used in legal terminology of these two concepts are informational self-determination and freedom of information.) In the areas below the horizontal line the main rules apply, above the line the exceptions. The overlapping (hatched) area represents the sphere of the personal data of public servants or public figures whose data relating to their public function is subject to the main rule of openness, instead of self-determination.

Székely's basic model can be criticized because data controlling relations of non-state (business or civil) organizations cannot be illustrated in it. Its consequently improved version is not comprised of three, partly overlapping circles or ellipses but of a ring of three long-shaped curved figures the neighbouring ends of which overlap each other. In *Heller's and Rényi's model* (reflecting a sociologial and mass communication approach) the private/public and accessible/non-accessible attributes of information are illustrated in a two-dimensional system of coordinates. Two of the resulting four fields can be considered as "natural", while the two others require an explanation, i.e. are exceptional.

The accessibility of data for anyone or only for determined persons, and their unaccessibility for others, are determined by numerous rights and interests. These rights and interests are realized in traditional secrecy categories, whose enforcement is associated with means and procedures according to the state of actual information technology. Significant nominative secrecy categories are: business secret, state secret, office secret, lawyer's secret, medical secret, banking secret, confessional secret etc. Some of these categories (e.g. medical secret) reinforce the main rules, others (e.g. state secret) constitute their significant exceptions, and can be represented in different fields of the model accordingly. The basic criteria of these categories are generally determined by laws and regulations; their actual application within this framework is determined by the responsible controllers of information systems. However, the personal/public dichotomy prevails when applying these secrecy categories as well.

It deserves to be noted that the above data controlling categories and main rules evolved long before the emergence of automated data processing and modern

telecommunications, but their pertinence has increased with the ever deeper penetration of ICT. In any era, applying modern information and communication technologies raises a number of theoretical and practical problems in the relationship between state and citizen, business and customer, or, in a general sense, the stronger and weaker party from an informational point of view. One main branch of problems originates in the change of the informational limits of privacy, i.e. the concentration of informational power *as a factor controlling the individual*. The other main branch originates in the change of the individual's informational status determining his social participation, i.e. the concentration of information power *as a monopoly of handling public information.*

While the historical achievement of informational self-determination can be described in terms of states of equilibrium and periods of disequilibrium induced by new ICT, the theoretical and practical realization of freedom of information can be interpreted as an evolutionary process. [1.8.1]

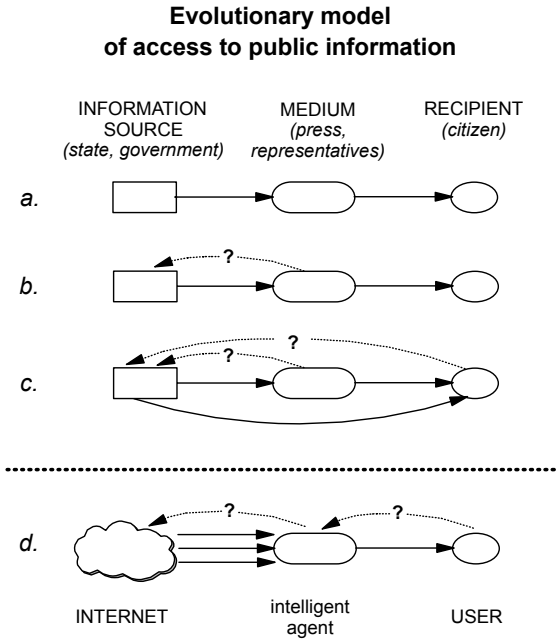


Figure 1.8.2

Stage *a.* can be considered a simplified information model of *representative democracy:* the representative (and also the media, in theory representing the citizen) has the right to transmit information from the source to the recipient. Stage *b.* is the model of *freedom of the press:* the medium has an exclusive right not just to

transmit but also to demand information. Stage *c.* illustrates the model of actual *freedom of information* where recipients have the right to demand and receive information directly from the source, with the exclusion of the medium. (Transmitting information can be performed with or without modern ICT.)

Freedom of information, together with the widespread use of the internet, has raised the illusion of *disintermediation.* Getting rid of the medium is an illusion not only because in the global network one can find (lawfully) only such public information which has been *posted* by someone in order to provide public access. The other reason is that internet users face two basic problems: the quantity and quality of information. One of the reintermediatory solutions of these problems is the use of an *intelligent agent,* as illustrated at stage *d.,* where the agent containing a personalized knowledge base, pre-selects, or "pre-digests" the information requested by the recipient. (It should be noted however, that such an agent used for facilitating access to public information can be a most effective, at present hardly controllable tool for manipulating its user.) Here "internet" naturally means only a virtual information source, there is always a real source behind it; the change of the actors and roles are indicated by a dotted line.

1.8.2. Definitions

The uniform use of terminology relating to the above is a general requirement in the field of modern information and telecommunication services nowadays. (In Hungary and in other countries even today high quality works are written using incorrect/outdated terms, and in circles of technical intelligentsia the twin concepts of data protection and data security serve as a root of misunderstanding.) The short definition of the most important terms are given below, according to internationally accepted terminology.

Data protection: the totality of principles, rules, procedures, data controlling means and methods guaranteeing the limitation of collecting, processing and use of personal data, and the protection of the persons concerned. (By definition, data protection can be interpreted only in the case of personal data.)

Data security: as interpreted here, is the technological and organizational system for preventing unauthorized access to, alteration of, and destruction of data.

(Data security can be interpreted in the case of both personal and non-personal data.)

In short: data protection is protection of the *data subject,* while data security is that of the data itself.

Personal data: any data relating to an identifiable natural person and any conclusion drawn form such data with respect to him/her. The data preserves its personal nature as long as its relation to the data subject can be (re)established.

Data of public interest: any data under processing by an authority performing state or local self-government functions or other public duties that does not fall under the category of personal data. (The determining element in the public-interest nature of data is therefore neither its declaration, nor the data controller's form of ownership, but public function.)

Data subject: the person concerned to whom data can be related.

Data controlling: recording, storage, processing, use, transfer and disclosure of personal data, irrespective of the means applied.

Data processing: performing data controlling operations or technical tasks, irrespective of the means and methods applied.

Data controller: the natural or legal person or organization without legal personality which determines the purpose of data controlling, makes and implements decisions relating to it, and is entitled to authorize a data processor to implement them.

Data processor: the natural or legal person or organization without legal personality which processes personal data on behalf of the data controller. (Data controller and data processor are therefore independent persons or organizations; at present the typical example of their relationship is *outsourcing*. The responsible party for the legality of data controlling and processing is the data controller.)

1.8.3. The basic principles of data protection

The next level of implementing the main rule and exceptions of informational self-determination is constituted by internationally accepted, formulated data

protection principles. Below, the principles are introduced in abbreviated form, following the categorization of the OECD Data Protection Guidelines [1.8.2].

1. Collection limitation principle

Personal data should be obtained only by lawful and fair means, and with the knowledge and consent of the data subject.

2. Data quality principle

Data, in accordance with the purpose of data controlling, should be accurate, complete and kept up to date.

3. Purpose specification principle

Personal data can be controlled only for pre-determined purposes, and only to the extent and duration necessary for achieving those purposes.

4. Use limitation principle

Data should be used only with the consent of the data subject or by the authority of law.

5. Security safeguards principle

Data should be protected, according to the state of available technologies, by reasonable security safeguards against loss, unauthorized access, modification, destruction or disclosure.

6. Openness principle

The existence, place and purpose of data controlling, the identity of the data controller, and the data controlling policy should be publicly available.

7. Individual participation principle

Data subjects should have the right to know, and where appropriate, rectify, complete or erase the data relating to him/her.

8. Accountability principle

The data controller should be accountable for complying with the above principles, and he/she should be able to prove the lawfulness of his/her data controlling.

1.8.4. National and international law

The following level of data protection is constituted by the system of conditions and requirements defined in international conventions, guidelines and other documents. The most important international documents are: the OECD Data Protection Guidelines, the Data Protection Convention of the Council of Europe [1.8.3], and the Data Protection Directive of the European Union [1.8.4].

The 1980 OECD Guidelines and the 1981 Council of Europe Convention were prepared in parallel. Both contain the basic principles of data protection, however, while the OECD Guidelines focus on the necessity (and legal guarantees) of transborder data flow, the aim of the CoE convention is to ensure information rights in transborder data flow. A further difference is that the adoption of the Guidelines is recommended only, while compliance with the Convention is obligatory for its parties.

The EU Directive, adopted after a five-year debate, determines the common, detailed rules of data controlling which the EU member countries were obligated to adopt into their national laws. The deadline of implementation was October 1998. Countries where data protection laws and practices were implemented earlier had to harmonize them with the provisions of the Directive, while newly adopted laws in both member and accession countries have been prepared already in the spirit of the Directive.

As Colin Bennett, Canadian author and professor of political sciences, discovered as early as the 1980's, data controlling led not only to technological convergence but "policy convergence" as well. However, important practical consequences derive from the fact that while the CoE convention requires an *equivalent level of protection* in the member countries, the EU directive requires only an *adequate level of protection*. Personal data can be transferred to so-called third countries (outside of the convention or the EU) without restrictions, e.g. through automated information and telecommunication systems, if the third country meets the required level of protection. Equivalent protection requires identical regulation and practice, while adequate protection can be realized in different regulatory environment and by applying alternative methods and means.

The problems of adequate protection appear strikingly in the debate between the EU and the United States. The differences between the European and the

American model are summarized in *Table 1.* The European model is represented by the West-European countries and the developed newly democratic countries, while the American model is followed, with minor differences, by the USA, Australia, New Zealand, and – earlier – Canada.

	European model	American model
Sector:	private + public	public
Processing:	automated + manual	automated
Scope:	general	mosaic-like
Supervisor:	Yes	No

Table 1.8.1. Legal regulation of data protection

In the case of a country regarded as non-adequate, restrictions have to be imposed on transborder flows of personal data, and in the era of globalization of information flows this can have significant political and economic consequences. At present the competent bodies of the EU, after a series of contradictory resolutions, regard the activity of US data controllers who voluntary follow the so-called Safe Harbor Principles, as being of an adequate level of protection. However, the number of such data controllers is low and there are no sanctions for breaching this set of codes. Hungary as the second non-EU country after Switzerland, officially received the status of "adequate" in 2000.

Data controlling conditions defined at the international level, however, are also to be directly applied in designing and operating data processing systems. For example, the three criteria of the data subject's consent as defined by the EU (i.e. freely given, expressed and informed) should be met in web-based online data processing systems too. Similarly, the implementation of the individual participation principle (for example, the separability and retrievability of transactions relating to individuals) requires a carefully designed identification system of data subjects in databases.

Internal (national) law is the following level of data protection. The Hungarian legislation and regulation is up to date; it follows the European traditions. Its special feature is that both data protection and freedom of information is regulated in a single Act. The legal hierarchy ranges from the Constitution through the framework-like Data Protection and Freedom of Information Act [1.8.5] to sector-specific Acts and decrees.

The implementation of the national data protection law is controlled by independent supervisory authorities. Regarding their competence and legitimacy these are of different types: there exist bodies like the French CNIL (Commission Nationale de l'Informatique et des Libertés) or unipersonal institutions like the British Information Commissioner (formerly Registrar); they can be elected by the parliament (Germany) or nominated by the government (The Netherlands); their competence can be authority-like, court-like or ombudsman-like. The Hungarian DP-FOI Commissioner, together with the two other parliamentary commissioners, was elected the first time in 1995. He has an ombudsman-like competence extending to both data protection and freedom of information. His core activity is investigating complaints but he is also entitled to start investigations *ex officio* in both public and private sectors. His investigations result in recommendations which are not binding in the legal sense, but their adoption rate is high. He gives opinions on draft legislation relating to data controlling, and supervises the grounds of state and office secret classifications. His competence in secrecy supervision is like that of an authority, his calls for declassification are binding. [1.8.6]

1.8.5. Privacy Enhancing Technologies (PET)

The various information and communication technologies called Privacy Enhancing Technologies (PET), representing the technological level of data protection, were developed in order to protect not only data but *data subjects* as well. The purpose of applying PETs in a broader sense is to reduce the invasion of privacy, namely to enable the use of advantageous new technologies without further loss of privacy, or at least reducing the loss. PETs, if properly used, always protect the weaker party (typically the data subject himself/herself) against the stronger party having excessive informational power.

PETs can be classified by several systems of categorization:

(a) According to Burkert's categorization [1.8.7] there are

- subject-oriented
- object-oriented
- transaction-oriented, and
- system-oriented technologies.

In the first case the technology is oriented towards the data subject (e.g. ensuring anonymity for card holders), in the second case towards the means (e.g. anonymous digital cash), in the third case the purpose is to delete the tracks of the transactions, and in the fourth case several technologies are integrated into a system.

(b) From another point of view one can distinguish

- technologies enhancing the security of existing systems,
- new data storage and retrieval technologies, and
- transaction-based technologies.

(c) PETs can be categorized according to which data protection principle they help to implement.

(d) Finally, one can distinguish

- technology-based, and
- human interaction based

PETs.

It is important to note that the same set of technologies can be classified according to different categorizations. Some examples of applying PETs:

Bioscrypt is composed of two starting elements: a biometric element (e.g. digitized fingerprint) and a non-biometric element (e.g. key, PIN or pointer) by "biometric encryption". The condition of using bioscrypt-based applications is re-producing the biometric element, i.e. the presence of the data subject. Such applications include the "anonymous database" where identifying data and the relating personal data are dispersed in the fields of a database in a quasi random way, and the link between the related data is stored in a pointer which is encrypted in a bioscrypt. According to categorization (b), this is a new data storage and retrieval technology, which, according to (c), helps implement the purpose specification principle. Bioscrypt can be used to hide the content and the identity of parties of interpersonal communication in networks; here a common bioscrypt of the parties is needed which can be decrypted by the fingerprint of any of the parties, thus retrieving a symmetric key hidden in the bioscrypt. It can also be used in electronic commerce applications: this represents a three-party (Bank, Customer, Business) communication, again using a common bioscrypt, but here the Bank, instead of a

fingerprint or other biometric element, uses a unique artificial pattern of required complexity to produce the common bioscrypt.

Platform for Privacy Preferences (P3P). This is an internet-based technology developed for English language e-commerce applications which converts the flow of personal data between the remote data controller and the data subject into a standardized bargaining process. It is a human interaction based technology which at the same time helps implement the individual participation principle (the data controller gets hold of the data of the subject with the latter's knowledge and control).

Anonymous remailer. Electronic mailing services which hide not only the content of the message but also its existence, time, frequency and the identity of the communicating parties from unauthorized observers. Their advanced (Mixmaster type) versions apply delaying and random forwarding of the messages, filtering out of identical messages, linking of a chain of remailers, as well as standard size and layered encrypted format.

Digital pseudonyms. Several PET applications offer co-called *nyms* to users of internet services. These digital pseudonyms can be used as constant aliases in communication with certain data controllers, or as non-recurrent identifiers.

Authentication without identification. A basic issue surrounding PETs developed for financial applications is the problem of "authentication without identification". Both theoretical research and experience gained from existing systems prove that the bank as data controller can prevent itself from being informed of the totality of its customers' transactions, and this, paradoxically, reinforces the level of data security of the system, and it can also guarantee the authenticity of the transactions. One of the two theoretical solutions of this problem is to cut the transaction into pieces in a way that authentication can be followed through the whole transaction while identification appears only between two respective points of the transaction, and at each point a one-way (i.e. hardly decryptable) conversion is performed on the customer's identifier. The other solution is performing repeated, trivial transactions by using non-recurrent digital pseudonyms.

1.8.6. Data protection expertise and tasks

Technological convergence and the spread of online services make it necessary to adapt the practical forms of implementing data protection requirements to the changing technological, legal and organizational environment. Special attention should be paid to the interpretation of the relatedness between data and data subject, the interpretation of the re-establishability of this relatedness; as well as satisfying the conditions for linking databases and the conditions of the online consent of the data subject; similarly, the conditions of transfer (especially online transfer) of data, the implementation methods of the right to inspect and erase data, and the relation between and functions of data controller and data processor.

In order to implement data protection requirements in both static and changing environments an expertise including knowledge in information science and technology in the broad sense, as well as law and social sciences, is needed. Data protection requirements and organization-level procedural rules in data controlling systems are suitably included in internal data protection regulations. An expedient method of checking data controlling systems from a data protection point of view is *data protection auditing.* This represents a coherent system of examining the data controlling processes of an organization, the methods and technologies applied, and the products and services containing personal data. (Data protection auditing is not equal to *data security* auditing; however, some results of the latter constitute parts of the areas examined in data protection auditing.) Preliminary data protection analysis is recommended when designing and realizing new personal data controlling systems. The data protection status should be checked regularly.

References

[1.8.1] Székely, I.: *Az adatvédelem és az információszabadság filozófiai, jogi, szociológiai és informatikai aspektusai. Kandidátusi értekezés.* [Philosophical, legal, sociological and informatical aspects of data protection and freedom of information. Doctoral thesis.], Budapest 1994.

[1.8.2] *Guidelines on the Protection of Privacy and Transborder Flow of Personal Data.* Organisation for Economic Co-operation and Development (OECD), Paris 1980.

[1.8.3] *Convention for the protection of individuals with regard to automatic processing of personal data.* Council of Europe, Strasbourg, 28 January, 1981. *European Treaty Series* No. 108.

[1.8.4] Directive 95/46/EC of the European Parliament and of the Council on the protection of individuals with regard to the processing of personal data and on the free movement of such data. *Official Journal of the European Communities* No. L 281/31, 23.11.1995

[1.8.5] Act No. LXIII of 1992 on protection of personal data and disclosure of data of public interest, Hungary.

[1.8.6] *Annual Reports of the Parliamentary Commissioner for Data Protection and Freedom of Information* 1995–1997, 1998, 1999, 2000 Budapest.

[1.8.7] Burkert, H.: Privacy-Enhancing Technologies: Typology, Critique, Vision, In: Agre, P.E. – Rotenberg, M. (eds.): *Technology and Privacy: The New Landscape.*

MIT Press, 1997.

1.9. Security

Tibor Nemetz dr., author

István Vajda dr., reviewer

1.9.1. Definitions

Recording knowledge and ideas has been important to mankind since prehistoric times: pictures, ideograms, symbols, and later writing were used to achieve this goal. Writing had a dual purpose: it allowed thoughts to be reproduced at a later time, and could be used to convey these thoughts to others far from the author (in space or time). As writing evolved, the need soon arose to keep unauthorized persons from gaining access to the recorded information, even if that person was able to get hold of the entire written text. This need was primarily felt by those in power, secular or ecclesiastic, and was self-evident and fundamental to warlords and diplomats. They practically monopolized both the theory and the practice.

Initially, the secrecy of information was ensured using physical means, hiding information in carrier media consisting of natural or artificially constructed text. The science of sending completely hidden confidential messages (e.g. written using invisible ink) is called **steganography**, Greek for “concealed writing”. As character-based writing evolved, soon another easier, yet secure way was found for hiding information: the method of substitution. One of the first character codes was used by Jews before Christ. Their method was to replace the first letter of their alphabet with the last, the second with the last but one, etc. A method that gained more widespread use was invented by Caesar, emperor of Rome, who substituted each letter of the message by the fourth letter following it in the 25-letter Latin alphabet, maintaining the spaces. This process, known as **Caesar cipher**, today means the substitution of each message letter by the one c positions after it in the fixed alphabet. Accordingly, specifying c is equivalent to specifying the key. Once c is known, enciphering the message and deciphering it from the enciphered message is easy.

At this point, let us introduce and define some basic terms and concepts.

Clear text: the written message to be encrypted.

Encryption: the process used to make written text unintelligible to unauthorized persons.

Encrypted text: the new text resulting from encrypting the clear text.

Cryptography in a broad sense covers the research and application of all processes, algorithms, and security policies that are aimed at hiding written information from unauthorized persons. In a narrower sense, the term cryptography is limited to the science of encryption algorithms. This more limited scope means the field of mathematics, encompassing several areas of that discipline.

Cryptographic algorithms are specific encryption transformations. They assign the finite series $x \in X^{\infty}$ of a finite input alphabet X to the finite series $y \in Y^{\infty}$ of a finite Y output alphabet. This assignment is implemented by a **one-parameter family of codes** as defined in information theory. During every encryption session, one of these parameters must be chosen and the corresponding code used. The selected parameter is called the key, and is selected from the set of possible parameters called the **keyspace**.

Key management is the policy used to select the key and convey it to the partner in a secure manner to ensure confidential communication. It also covers the methods used for the safe keeping of the key.

Cryptanalysis: comprises all activities aimed at restoring the clear text from an encrypted text.

Cryptology combines cryptography and cryptanalysis. It also includes the physical means and intelligence-related issues of unauthorized decryption (“cracking” the key”). The Caesar cipher in its general form allows as many keys as there are letters in the alphabet (including the trivial case of **c=0**, i.e. no letter shift). An attacker who knows that the encryption process used was the Caesar cipher can decipher the original message by exhaustive trial (brute force). All the attacker needs do is write down all possible “shifted” text strings using all possible values of c, and select the one that makes sense. The definite article suggests that there is only one deciphered text that makes sense, and this is in fact true if the text is sufficiently long.

Defining the term “sufficiently long” allows us to introduce the term **unicity point**, a key concept of algorithmic encryption. The unicity point of an encryption algorithm is the minimum length (called the unicity distance) of the symbol string of ciphertext that an attacker must intercept in order to determine the plain text or the key. If no such distance exists, its value is considered to be infinite. Since the distance depends on the clear text, it is a statistical variable. The unicity point is frequently defined as the expected value of the unicity distance, although it is difficult to determine its exact value.

For European languages, the unicity point of the Caesar algorithm is 4–5 letters. Rather than providing a formal proof, we recommend an “empirical” verification of the truth of this statement. Select a block of 4–5 letters from a book or newspaper, and try to complete it to form an intelligible text by prefixing and/or appending additional letters, e.g. by exhaustive trial. You will find that in the majority of cases there will be only one usable solution.

1.9.2. Classical encryption procedures

A variant of the Caesar cipher that is more difficult to decipher is **simple substitution**. This process replaces each letter of the input alphabet of the clear text by another letter of the output alphabet, always using the same output letter for an input letter and assigning each input letter to a different output letter. Since the number of possible output alphabets is extremely high, one is tempted to believe that the key space is infinite. However, most of those possibilities are equivalent due to the following theorem:

Theorem: Any arbitrary substitution using any arbitrary output alphabet is equivalent to a substitution using an output alphabet = input alphabet, because one can only be decrypted if and only if the other one can also be decrypted.

This theorem can be used to **verify** the **security** of simple substitution, and to determine the number of all possible different codes for an alphabet of a given size. This number is equal to the number of possible letter assignments. These assignments are permutations of the alphabet, so for an alphabet size **k**, the number of different codes will be **k!**., Even for the Latin alphabet where k=25, this number is too large for finding the code by exhaustive trial, giving rise to the long-held

misconception that, in contrast to the Caesar cipher, simple substitution is unbreakable. Experience teaches us, however, that even grade schoolers are able decrypt a text consisting of a few hundred letters. Decryption can make use of the *statistical properties* of the language. In all languages, some letters occur more frequently than others. Accordingly, letters in the cipher text substituted for more frequent letters of the clear text will be more frequent than those substituted for less frequent letters. The same applies to letter pairs (bigraphs) and letter triplets (trigraphs). Experienced cryptographers are able to decrypt even a very short encrypted text. It is an empirical fact that for English, the unicity point for simple substitution is at 20–25 letters. (This is, of course, only enough to determine the clear text, but not the transformation itself.)

The fact of decryptability weakness was already known in ancient times. Therefore, simple substitution was superseded by a more advanced technique called **transposition**. The idea here is to divide the clear text into contiguous blocks of identical size, and then replace each letter in a specific position within the block by another letter at another position, applying the procedure for all blocks. This method is not vulnerable to simple letter frequency analysis, since it maintains letter frequencies. However, the higher level analysis of letter pairs can still be applied to this method. This approach leads us to the area of *statistical hypothesis tests*. It follows the strategy of small steps. Rather than attempting to determine the entire permutation used for scrambling, it only tries to find the letter positions of the clear text that follow the letter positions in the cipher text.

By the end of the Middle Ages, it became apparent that the reason why simple substitution can be decrypted so easily is that it only uses a single substitution alphabet. This realization lead to the use of **polyalphabetic procedures**. One such procedure was developed by Blaise de Vigenère (1523-1596). His comprehensive work encompasses the entire science of cryptology of his age. He developed several polyalphabetic procedures that used an addition operation defined for letters of the alphabet, which makes him a forerunner of the theory of finite groups. He didn’t live to see his methods used; their importance was only realized 200 years later.

The Vigenère cipher named after him is not only of historical importance: it led to the invention of an encryption algorithm that theory proved to be unbreakable. The addition used in the procedure is extremely simple. The method involves arranging

the letters of the alphabet in alphabetical order, and sequentially numbering them starting from 0. Two letters are added by adding their respective sequence numbers. If the resulting value is less than the number of letters in the alphabet, the result is the letter with the sequence number obtained. If the sum of the two sequence numbers is more than the number of letters in the alphabet, the number of letters in the alphabet is deducted from the sum. The result will again be less than the number of letters in the alphabet, so the alphabet will have a letter with this sequence number. In this case, the result is the letter with this sequence number. This procedure is obviously possible for any two letters of the alphabet. Also, equations of the type

a specific letter + an unknown letter = another specific letter

can be solved for any two letters, and has a unique solution. Now that you know the theory, try it in practice:

- Write down the clear text
- Choose a keyword
- Write the keyword above the clear text, repeating the keyword as needed
- Add each letter column. The resulting series of sums is the encrypted text. Obviously, text encrypted with this method can always be decrypted using the relationship below:

Unknown clear letter = cipher letter - keyword letter

This process is inherently vulnerable to attack. The main problem is that *the key length can be determined using a simple statistical test*. Once the length of the keyword is known, encrypted letters corresponding to the same keyword letter can be collected in a column, reducing the cipher to a series of monoalphabetic ciphers, each encrypted with a simple Caesar cipher substitution, the key of which is the corresponding keyword (and this is just one way of decrypting a Vigenère code!).

The Vigenère process can be cracked because the keyword length is finite, allowing the cipher text to be broken down into parts encrypted with a single code alphabet. This method can be thwarted by choosing a keyword at least as long as the clear text. If we could make sure that the cipher text and the clear text are statistically independent, i.e. that knowledge of the entire cipher text does not add to an attacker's knowledge about the clear text, this would render unauthorized decryption hopeless. This was first realized by Vernam, holder of the patent of the Vernam cipher or **one-time pad**. He modified Vigenère's process by choosing a very long

keyword and choosing each keyword letter randomly, proceeding letter by letter. The series of keys so generated was written on two writing pads, one for each party exchanging encrypted information. Pads had to be exchanged via a secure channel. Each pad was used only once, hence the name *one-time* pad. During WW2, the powers used various versions of the one-time pad to transmit sensitive information.

The requirement of random selection highlights a tenet of cryptography generally recognized as early as the 19th century.

Any cryptographic process must assume that the attacker has full knowledge of the parameterized code family (e.g. because he has bought it from an employee). The only remaining guarantee of maintaining secrecy is the selection of the key, so it must be chosen to represent maximum uncertainty for the attacker. This can be ensured by choosing a new key for each new clear text, and do so independently from the previous one, i.e. completely randomly. Methods for accomplishing this represent a major area of mathematics, *random number generation*.

The 17th century saw yet another generalization of the concept of simple substitution. This involved substituting letter groups (instead of letters), with the substituting sequence usually shorter than the substituted string. This method was discovered by Antoine Rossignol. He collected the typical expressions and names occurring in the communications of the royal court, and organized them in a **code book**, identifying each entry by a short string. Constructed in the second half of the 17th century, his code book called the Grand Chiffre defied all attempts at decryption until 1890. The early history of code books also had bearings on Hungarian history: as part of his foreign policy aimed at breaking Habsburg power, Louis XIV, the “Sun King” sent code books to Ferenc Rákóczi II, prince of Transsylvania, to support their covert communications.

Code books were the first step towards the commercialization of encryption. Shortly after Morse’s invention of the code named after him, Francis O. J. Smith, a lawyer, wrote a best selling code book for the transmission of secret messages on open wires using Morse’s telegraph. His code book demonstrated the close relationship between data compression and encryption.

In the wake of the industrial revolution, the growth in the number and size of encrypted messages made the mechanization of encryption a pressing necessity.

The forerunners of today's encryption machines appeared already in the second half of the 19th century. They all supported the use of polyalphabetic processes. The first encryption machines implemented different flavors of the Vigenère process. Examples for these early machines are the Wheatstone disc from 1867 and Baseries' encryption disc from 1891. These two latter methods of encryption were predominant until the beginning of World War II. One of the two most advanced solutions used code books, and the other used disc-based encryption machines where rotating the discs automatically generated the encryption keys. It took surprisingly long to realize that these methods do not ensure efficient confidentiality.

A typical example for the delusion of uncrackable encryption is the ENIGMA machine used by the Germans in World War II. This machine did in fact use a highly complex mathematical algorithm. Its operation was reconstructed by two Polish mathematicians, and decryption work was moved to England when Poland was overrun. By the end of 1942, the British were already able to read the Enigma messages of the Germans "on-line".

Of course, a number of other mechanical and electrical encryption machines were also used besides Enigma. The most popular encryption machines were the Hagelin machines that are still in use today (although in a modified form). Their decryption algorithms are also public.

Kahn's work [1.9.4.] provides a detailed description of the history of cryptography.

1.9.3. Shannon's theory

Radio telecommunication came into general use during World War II. Such communications were routinely intercepted. As a result, code breakers amassed a huge body of encrypted messages that had been created with the same algorithm, although mostly with different keys used for different messages. This is one of the reasons why code breakers prevailed over code makers at the beginning of World War II. The algorithms of several encryption machines and other methods thought for long to be unbreakable were successfully broken. The resulting disillusionment called for a theory that is able to investigate the *reliability of a cryptosystem with scientific rigor*. The precondition to this was to first model the communications environment in

which cryptography was operating. This task, born out of necessity, was solved by **Claude E. Shannon** during Wold War II and published between 1948–1949 [1.9.8.], [1.9.9.]. Shannon developed the **mathematical theory** of communication and the **communication theory of secrecy systems**, both in use to date. His theory was able to formally prove that one-time random transcoding, patented by Vernam, is a theoretically uncrackable encryption method if decryption is to be performed knowing the encrypted text only.

The mathematical models of communication and cryptography that C. E. Shannon has created based on his experiences as a code breaker in WWII are considered to be the foundation of **information theory**.

Figure 1.9.1 shows the block diagram of the model. The figure also shows the possibility of an attacker intercepting the information flow. This model also highlights the fact that even the most secure, theoretically uncrackable encryption algorithm may be decrypted with ease if a **carefully planned operation policy** is not introduced and fully enforced.

To demonstrate this, here is a negative example of a theoretically uncrackable one-time transcoding scheme that can be compromised if the operation policy is not enforced. Using the same transcoding row repeatedly is a gross mistake. Key repetition can be recognized using a simple statistical test (even manually). If an attacker detects key repetition, all he needs to do to get rid of the key – the guarantee of security – is to extract two encoded messages from each other, which cancels out the key. The difference text so obtained will be the difference of the two clear text messages, both of which can be decrypted (eventually with slight mistakes) using the method developed by *Kerckhoff* in the 19th century.

The mistake of key repetition was committed by several countries, including the Soviet Union. The reason was the significant cost of “pure” usage. Consider the cost of a real random number generator. Then add the cost of storing the generated keys securely and transmitting them to users through a secure channel, and that at a time of war. Several countries cut corners by using the same sequence of random keys repeatedly, applying a certain amount of shifting. According to the studies published under the title *Venona papers* in 1995, US intelligence set up a separate team for collecting and decrypting telegrams that used the same keys repeatedly. In

this case, the “auxiliary information” shown in the diagram of the model was the information acquired by intelligence that repeated keys were being used.

Auxiliary information may facilitate code breaking, or even render it superfluous. Electronically reading the text displayed on a computer’s monitor provides direct access to the clear text. Any “good” security policy should consider such and similar characteristics of the encryption environment. The discipline dealing with the application of cryptographic methods covers much more than just algorithmic issues.

1.9.4. Public encryption processes: DES, AES, IDEA

Cryptography has traditionally been used by the military, the intelligence agencies, and diplomacy. All of these organizations are typically centralized. They usually set up an intelligence center entrusted with the task to select the encryption algorithm, develop and enforce the security policy, and organize the continuous generation and distribution of encryption keys. The users of the application are well-defined communication end points that form a closed network.

Closed network: a communication network where all participants are a priori known and registered. Nobody can voluntarily enter the system as a new user. Only users first authenticated by the network are allowed to generate traffic. The explosive development of communication also created the need for even unknown persons to be able to communicate securely with each other, without third parties gaining knowledge of the content. Accordingly, closed networks are increasingly replaced by open networks.

Open network: a communication network that anybody can join in accordance with published rules.

The first communities wishing to communicate securely were groups of businessmen who knew each other. This was not a significant departure from classic applications. They could mutually agree on the encryption algorithm to be used, choose their encryption keys valid for a limited period, and exchange them personally among themselves. Their network was also closed, with the difference that extending the network was easier to administer and could be decided by mutual consent. Such an **open network** may also be centralized, like that of a company and its affiliates (or

customers). In these cases, the cryptographic algorithms themselves remained unchanged, and traditional methods could be used. However, due to their lack of specialist knowledge and their limited resources, they needed technical help in choosing the algorithms and developing their security policy. Implementing such networks was a matter of national interest, as first realized by the United States. An encryption standard was needed to guarantee secure communication for the users of the network. The first such standard was **DES** (Data Encryption Standard).

It was announced in 1976 by the *National Bureau of Standards* of the USA as a new Federal Information Processing Standard (FIPS 46). Its description can be found at the United States Patent and Trademark Office (see National Bureau of Standards [1.9.6.]). DES uses a 56-bit key to map clear text message blocks (of 8 bytes = 64 bits) to encrypted blocks of 64 bits. The block algorithm was developed by IBM using a Shannon-type scrambling transformation. It ensures that every bit of an output block is dependent on every bit of the input block. The standard also stipulates that within the United States, the algorithm can only be used in hardware implementations, and the US government prohibited the export of this hardware implementation. The algorithm itself was subjected to a security review every five years. The last such review occurred in 1994, when the final year of its usability was set to 1998. Despite this, a slightly modified version of DES is used to this day.

The series of publications investigating possible ways to break DES started with a presentation held by Martin Hellman in 1977. He stated that exhaustive trial, or brute force attack against DES was viable using a suitably constructed hardware. The decisive blow that led to the death of DES was delivered by Biham and Shamir, who gave a possible decryption process based on their new method called “differential cryptanalysis”. In 1994, Matsui devised another method called linear cryptanalysis that he successfully used to attack the DES cipher. As a first attempt at remedy, the suggestion was made to apply DES twice in a row, stating that this would render cracking practically impossible. This involved encrypting the message twice using two statistically independent keys, which effectively meant increasing the key length to 108 bits. However, Merkle and Hellman proved already in 1992 that if “simple DES” can be cracked, then an attack based on the “man in the middle” method allows cracking “double DES” as well.

The current most advanced version of DES is triple DES a.k.a. 3DES. This method uses two or three 58-bit keys. The message is encrypted in three steps, by first doing a DES encryption in normal DES mode with the first key, a DES decryption of the result using a second key, then a DES encryption of that result in normal DES mode with the first key (if two keys are used) or with a third key. Several custom machines are said to have been built that can decrypt 3DES.

Paradoxically, export restrictions caused a number of chip manufacturers to implement their own solutions and offer them commercially. Since several international standards of the finance industry include 3DES components, prospective buyers of 3DES chips are well advised to critically investigate the solution they want to implement. They are definitely advised against using software downloaded from the Internet. As the algorithm is public, it can be programmed without infringing copyright. A number of mathematical tricks can be used to decrease the processing time.

In the fall of 2001, the National Institute of Standards and Technology (NIST) announced a new US encryption standard called **Advanced Encryption Standard** (**AES**). The shortcomings of **DES** forced the decision of NIST to develop a successor of DES called **AES**. The Request for Proposals issued in September 1997 detailed the list of criteria to be satisfied by the AES algorithm. It was also declared that the submitted encryption algorithms would be made available on a worldwide, non-exclusive, royalty-free basis. The RFP criteria stated that the algorithm should:

- be block-based with a block size of 128 bytes
- provide the option to select a key size of 128, 196, or 256 bits
- be worldwide public and available for non-exclusive use without royalties
- resist all known decryption attacks
- be clear, logically structured, and relatively simple
- ensure fast encoding and decoding
- use as little memory as possible
- be suitable for implementation on a broad range of processors.

An encryption algorithm that satisfies these criteria is expected to provide a long-term solution for the secure protection of private data generated in the 20-25 years to come. From among the submitted solutions, 15 met all formal criteria. Held in Rome between March 22-23, 1999, round 2 of NIST left only five finalists.

The winner, announced on October 2, 2000, was the **RIJNDAEL algorithm**, named after its inventors, the **Belgian cryptographers Rijmen and Daemen**. The Rijndael algorithm fully meets all criteria described above. It uses several interesting techniques (like byte-level operations on a finite field of 256 elements). Its structure, implemented in several iterative steps, does not follow the structure of DES. Each iteration consists of three layers (a linear mixing layer, a nonlinear layer, and a key addition layer), each playing a different role (see [1.9.1]).

The Rijndael algorithm is simple, clear, and easily programmed in any programming language. Several flavors can be downloaded from the Internet, although this is definitely not recommended.

Already prior to the AES RFP, many solutions were proposed for replacing DES. The one in most widespread use is the **IDEA cipher** proposed in 1990 by Lai and Massey as a Proposed Encryption Standard. In 1992, Lai published an improved variant of the IDEA cipher known as IDEA (Ideal Data Encryption Algorithm).

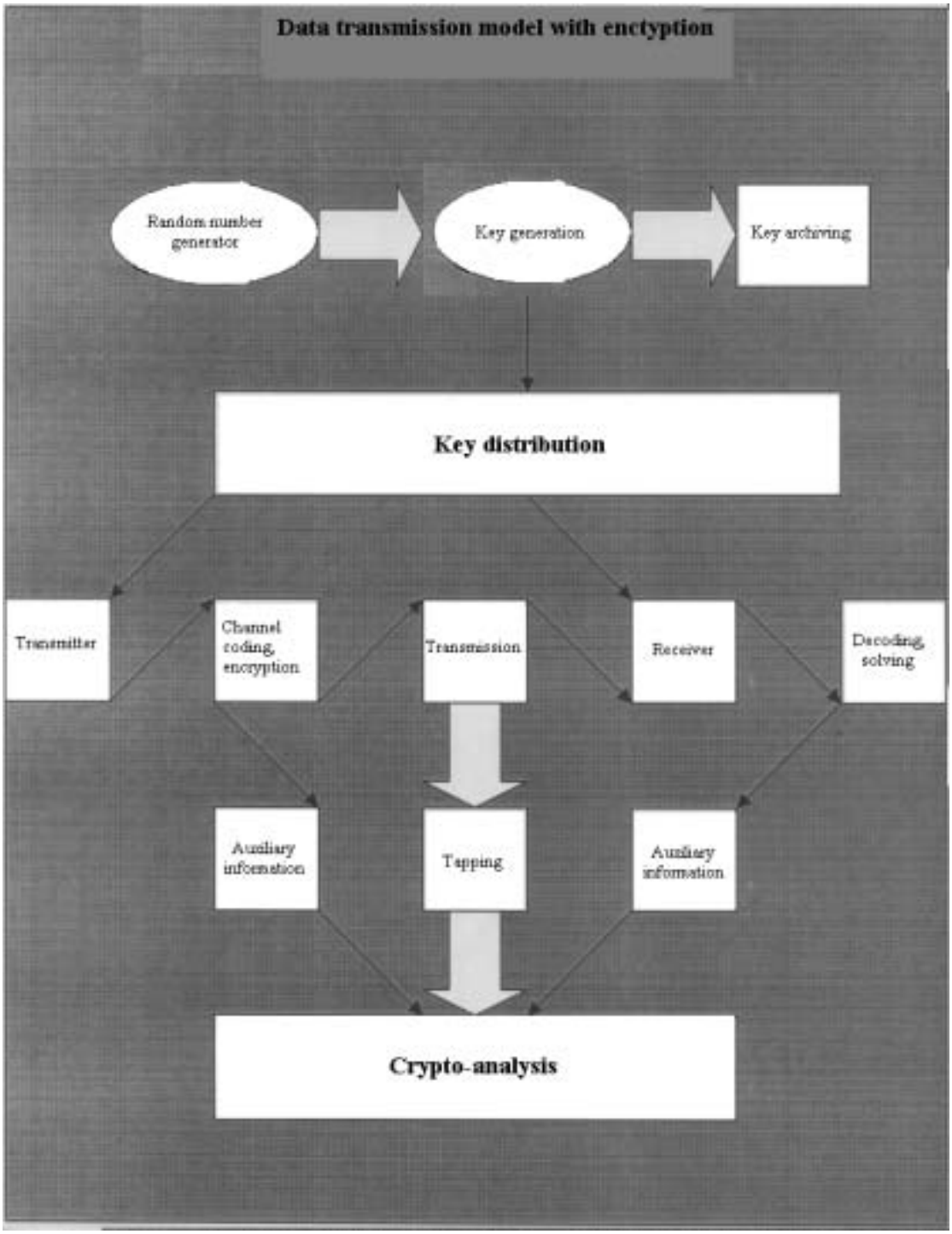
IDEA uses special combinations of carefully selected basic but sufficiently complex mathematical operations. It applies these operations in blocks of 16 bits on 64-bit blocks of clear text, using a 128-bit key. Within a block, every output bit depends on every input bit. The method can be proved to exhibit the confusion and diffusion properties required by Shannon's principles. The simplicity of the mathematical operations involved allows fast and simple implementation in both software and hardware. Some of these solutions are vulnerable to attacks, an issue to be kept in mind when shopping for a solution.

The IDEA process is patented in several countries. The European patent was granted on June 30, 1993, under patent number EP 0 482 154 B1. Any use of the algorithm in commercial products is subject to a license from ASCOM. Further information is available at http://www.szcom.ch/Web/systec/security/license/htm.

Classical encryption algorithms use the same key for encryption and decryption (although the decryption algorithm is not necessarily the encryption algorithm executed in reverse order). For this reason, such algorithms are referred to as **symmetric key algorithms**.

The first **speech scramblers** also relied on the methods of classic cryptography. These devices divided the unit interval of the time axis into many sub-

intervals of equal size, and permutated these sub-intervals. They also permutated small intervals of the amplitude axis subdivided in a similar manner. The current



practice is to first convert analog signals to digital, then encrypt the resulting digital sequences regarded as clear text.

1.9.5. Public key cryptography

We are witnessing the revolution of digital communication. Just as previous major changes in society resulted in specific cryptographic procedures adapted to the needs of the age, the electronic society is currently developing its own set of such procedures. One of the requirements of digital communication is for people unknown to each other to be able to exchange information in a secure manner. The technology foundations for this already exist. In a world of digital networks, security means satisfying two different types of requirements. Safeguarding privacy and data security requires different actions for each. Let us first define these two areas:

Data security: the protection of digital data from unauthorized modification, destruction, or disclosure.

Privacy: the protection given to information to prevent the unauthorized disclosure of private data, e.g. through access control. The tool for safeguarding privacy and data security is public key cryptography.

Public key cryptography is a cryptosystem where participants use a common encryption algorithm. It is used on public networks that anybody can join. The algorithm uses two unique keys. One is the public key that the participants publish together with their names; the other one is the private key that they keep secret. One of the keys is used for encryption, the other one for (authorized) decryption.

The common cryptographic algorithm makes encryption with the public key easy, but decrypting the message using the public key is practically impossible. However, the private key will quickly decrypt the message. Systems implementing this philosophy are collectively referred to as *public key cryptosystems*.

The concept was developed by *Hellman* and *Diffie*. The first and to this day the only reliable and technically viable implementation was developed by Rivest, Shamir, and Adleman in 1978 (see [1.9.7.]). Its mathematical foundation is *Fermat's theorem* from *number theory* and the extreme difficulty of factoring large numbers.

The **RSA algorithm** rests on the extreme complexity of solving modulo arithmetic equations that contain the unknown quantity in the exponent, making the technical implementation of an attack impossible, provided that the modulus is

sufficiently large. The term “sufficiently large” plays a key role here: while a key length of 40 bits was initially thought to be secure, keys shorter than 1024 bits are considered insecure by today’s standards. The public key is a pair consisting of integers (E, M). They are used to perform the encryption. First, specified size blocks of the clear text are converted to integers smaller than the modulus M, then this number is raised to the power E in modulus M. The encrypted message is the resulting number, or the sequence obtained by encoding it for the transmission channel. Similarly to the public key, the private key is also a pair of integers (D, M), where M is the same as above, while D is the decoding exponent chosen to ensure that raising the modulo M number (i.e. the encrypted message) to the power D yields the original message. In order to obtain a reliable algorithm, M is chosen as the product of two very large prime integers, and E is chosen randomly.

Since exponentiation in modular arithmetic is costly in term of time and money, hardware solutions are used almost exclusively. Several such chip implementations are commercially available. They mainly differ in the selection of the random prime or the random exponent, but the decisive difference is the method they use for performing large integer arithmetic. The first aspect is important for filtering out solutions using “wrong” primes that allow unauthorized decryption, while the second is important for ensuring processing speed. Some of the first applications of RSA are digital signature, access control, and message authentication.

A digital signature can be used to guarantee document authentication (with regard to the originator of the message), document integrity, and non-repudiation. A digital signature is expected to provide the same signature properties for electronic documents as those provided by a manual signature placed on a traditional paper document.

A **digital signature** is an encrypted digital character sequence which inherently guarantees with a very high probability that it was issued by its owner, and which

- can be stored on electronic media
- can be transmitted via an electronic channel
- is dependent on the signer
- is dependent on the text signed

- may include a “fingerprint” or “message digest” generated from the message sent
- may include a time stamp of the signature.

When long documents are signed digitally, a short string called the *digital fingerprint* is assigned to the document using a method that is easily performed and guarantees that different documents are assigned different fingerprints. Fingerprints can be used for identification. The mapping that solves this task is called a **hash function**. Hash functions are expected to:

- make it practically impossible to construct a new document that has the same fingerprint as an existing document
- make it practically impossible to construct two documents that have the same fingerprint
- change significantly if even a single bit of the document is changed
- be easy to compute
- behave randomly for sufficiently long input documents
- treat inputs of 5-6 or more 32-bit characters as “sufficiently long”.

While several existing procedures meet the above criteria, the most popular one is the **Secure Hash Algorithm** (RSA). The input of the algorithm is an arbitrary document of arbitrary length (not longer than 2^{64} bits), and its *output* is a *160-bit string* (*hash value* or *message digest*.)

The standard of the SHA function is described in document No. FIPS-180 of the series of *Federal Information Processing Standard*s. The standard was slightly modified in 1996. The new version. called SHA-1, has been publishes as **FIPS PUB 180-1**. The algorithm can be implemented using many different computational procedures, but all must yield the same message digest if applied to the same input sequence. The standard specifies the following three strings for verifying compliance:

We note here that **MD4** and its modified version **MD5** are also considered acceptable. Another popular hash algorithm is **32-bit CRC** proposed by the CCITT (ANSI standard X3.66).

MD5 is subject to royalty payment, while the use of SHA-1 and CRC in proprietary software is royalty-free. Implementations for all of them exist and can be downloaded from the Internet, but there is no guarantee that they are correct or virus-free. Download pages typically warn visitors that the export of cryptographic software

is subject to the Export Administration Regulations (EAR) codified at 15 CFR Parts 768-799.

With e-mail messaging becoming predominant, the need arose to be able to send electronic messages that have the same legal effect as manually signed paper documents (e.g. offers, acceptance of offers, confirmation of performance, etc.). For this to work, however, the parties must be able to authenticate the sender of the message and verify that its contents have not been changed after posting. Digital signature satisfy these requirements. In order to ensure the legal acceptance of digital signatures, several countries have introduced digital signature legislation. The first European country to do so was Germany in the fall of 1997.

In June 2001, the Hungarian Parliament, recognizing the importance of digital communication, passed the **Digital Signature Act**, laying the foundations for authentic electronic statements and providing the legal framework for electronic communication in the fields of commerce, public administration, and other scenarios of the information society.

Some key definitions of technical terms from the Act:

Electronic document: any data that can be interpreted by an electronic device.

Electronic signature: electronic data or an electronic document inseparably connected or logically linked to the electronic document with the purpose of identification.

Electronic deed: an electronic document containing a statement or the acceptance of a statement, or the acceptance of a statement as obligatory.

Time stamp: a logical confirmation inseparably assigned or logically linked to the deed or document, containing the date and time of the application of the time stamp and technically linked to the document content in such manner that any modification to the document made after the application of the time stamp can be readily detected.

The time stamp is normally placed at the beginning of the document before the message digest is generated. The time stamp may also include *data that identify the signer*, e.g. passport number, date and place of birth, etc. The next step is the

generation of the message digest, so it becomes also dependent on the time stamp. The signature can be completed by using public key encryption. The corresponding procedure is performed by the signer using his/her private key.

In order to perform *signature checking*, the recipient first decrypts the signature with the signer's public key (changing the signature back into a message digest), then recalculates the message digest and compares the result with the message digest found in the signature. It the two match, the signature is verified as truly signed by the sender. However, this check can only prove that the signer is identical with the person who has placed the public key in the central key repository. So it should also be verified that the signer is in fact the person he purports to be. This is done by an **Authentication Authority**, a trusted third party with much the same role as a *traditional notary public*. In Hungary, Authentication Authorities are licensed by the Telecommunications Directorate in accordance with strict legal regulations. These include extremely rigorous security requirements that the infrastructure of an Authentication Authority must satisfy in order to perform its tasks. Beside secure encryption methods, these requirements also prescribe IT requirements like the use of *reliable firewalls*, and also cover aspects of the physical environment, e.g. contingency plans, personnel, etc. The Hungarian term for infrastructures that meet international conditions and regulations is **PKI**, borrowed from the English **Public Key Infrastructure**. The Act also defines some fundamental legal terms:

1. **Signature-generating data:** a unique data item (typically the cryptographic private key) used by the signer to generate the electronic signature.
2. **Signature-validating data:** a unique data item (typically the cryptographic public key) used by the recipient of the electronic deed or document for validating the electronic signature.
3. **Signature-generating device:** the software of hardware used by the signer to generate the electronic signature, using the signature-generating data.
4. **Signer:** the natural person linked with the signature-validating data as certified by the registry of signature-validating data published by the authentication authority.

Although by 2002, most European countries have passed their digital signature acts, their harmonization and fine-tuning is still on the agenda. Under the auspices of the Information and Communication Technology Standards Board, the European industrial and standards organizations have launched the European

Electronic Signature Standardization Initiative (EESSI). The goal of the EESSI is to perform a comprehensive analysis of future standardization needs in conjunction with the implementation of the EU Directive on digital signatures, with special regard to the business environment. The law makes it possible to certify that a deal has in fact been closed

- with the specified content
- at the place and date indicated
- with a signature that can be authenticated, and
- is linked to a single person.

  Accordingly, a digital deed is only valid if it meets the following conditions:

- non-repudiation: authorship and content can not be repudiated
- date stamp: the creation date can be verified
- integrity: neither the author nor any other party can modify it undetected
- confidentiality: unauthorized third parties are unable to read it
- authenticity: the person who created it can be verified to be who it purports to be.

1.9.6. Cryptographic protocols

An important pro of PKI is that the conditions for quickly and automatically connecting the two endpoints are available. A part of this task is performed by the **cryptographic protocols**. In addition to setting up the connection, these protocols also implement policies for ensuring that the algorithms used in a specific application provide the required level of confidentiality or authenticity. The two fundamentally different approaches that have proved to be efficient in an open network environment are network-level authentication and application-level authentication. Each operates at a different network level, which largely determines and limits their respective characteristics and features.

As an illustration of the protocols used in practice, a brief description of the **Secure Socket Layer** (SSL) protocol is given below. It is used by the server to authenticate itself to the client. The server responds to the challenge (“Who are you?”) of the client along the following lines: “I am XY, here is the certificate of my public key, and I can encrypt using the following algorithms: (list).” The client then generates a random master key and encrypts it with the server’s public key (RSA

application). The server uses its private key to decode the master key (RSA), then sends the client a message encrypted with the master key in order to authenticate itself. Now the client derives the connection key from the master key (MDS application), which is also performed by the server. Finally, the client authenticates itself. (The new version of SSL is called TLS.)

The **S-HTTP** protocol includes functionality for message authentication, digital signature, key generation, and encryption. As with HTTP, commands for cryptographic functions are issued between parentheses, including also the necessary parameters. The two parties negotiate the possible connection parameters (key generation, encryption algorithm and its mode of operation, hash function, method used with digital signatures, etc.), then embed the response fields in the HTTP file in accordance with conventions.

Using an interesting form of public-key cryptographic protocols, a party (the prover) can prove to another party (the verifier) that it knows a certain secret, without passing on any part of the secret. These protocols are called **zero-knowledge protocols**.

The system administrators of systems and networks providing cryptographic services must *come to grips with a number of tasks during their daily work*. Here are a few of these, for the sake of illustration. Prior to organizing a complex protection system, managers should conduct a rigorous site survey that covers:

- the encryption devices (the technology) that can be purchased,
- the interests as well as the financial and technical options of potential attackers,
- the qualifications of their own (service) staff.

They should formulate the basic principles (the security and privacy philosophy) that specialists of the different areas will use as the foundation of their specific implementation proposals.

The core task of the system designer is to check the costs and (financial) risks involved in implementing different levels of security, and to check how much it would cost an attacker to penetrate the system. Making the appropriate decisions is, however, outside the competence of the security expert.

The task of the system administrator is to enforce the accepted data management policies. This includes protection against attacks from within (by their own staff). The requirement of continuous control spawns a number of system administration tasks like logging, random uploads, substitution of staff members, prevention of unauthorized help, etc.

In the case of databases, the law on privacy and data security must be adapted to local conditions, and a stringent in-house policy must be developed for the management of confidential data.

If the decision is made to deploy a new encryption algorithm, it should be tested with the assumption that, although unable to access the encryption keys used, the potential attacker is fully informed about the encryption process and is able to compromise it.

EMI (electromagnetic interference) is especially important where information is transmitted electronically. The interception of electromagnetically radiated information should be prevented.

References:

[1.9.1.] J. Daemen, V. Rijmen: *The design of Rijndael*, Springer, 2002

[1.9.2.] Diffie, W. (1988): *The first ten years of public-key cryptography*, Proc. of the IEEE, 76. 560-577.

[1.9.3.] Diffle, W. – Hellman, M.E. (1977): *New directions in cryptography*, IEEE Trans. on Info. Theory, IT-22, 644-654.

[1.9.4.] Kahn, D. (1967): *The codebreakers*, MacMillan, New York

[1.9.5.] MEH (1996): *Informatikai rendszerek biztonsági követelményei*, Miniszterelnöki Hivatal, Informatikai Koordinációs Iroda (in English: *Security Requirements for IT Systems*, IT Coordination Bureau of the The Prime Minister's Office in Hungary)

[1.9.6.] National Bureau of Standards (1977): *Data Encryption Standard*, Washington, D.C.

[1.9.7.] Rivest, R. L. – Shamir, A. – Adleman, L. (1978J: *A method for obtaining digital signatures and public-key cryptosystems*, Comm ACM 21, 120-126.

[1.9.8.] Shannon, C. E. (1949): *Communication Theory of Secrecy Systems*, Bell Syst. Techn., 28,656-715.

[1.9.9.] Shannon, C. E. (1951): *Prediction and Entropy of Printed English*, Bell Syst. Techn. J., 30, 50-64.

Translated by Szombathy Csaba

1.10. Graph Theory and its Applications

Péter Laborczi dr., András Recski dr., authors

György Lajtha dr., reviewer

Graphs are efficiently manageable mathematical representations of telecommunication networks: the vertices of a graph correspond to routers or switches while edges represent the cables, radio and fiber links that connect the routers. In this chapter terms and methods are summarized that are essential for modeling telecommunication networks.

1.10.1. Introduction

In the beginning physical transmission medium has been directly used to transmit information, consequently it was easy to handle. The growing need to transfer information requires networks that are able to carry data of complex structure (picture and voice) efficiently. Since Internet Protocol (IP) is the most widespread application, which enables different paths for different packages, an efficient traffic management model is of great importance. MultiProtocol Label Switching (MPLS) that satisfies the above requirements in a simple way has been worked out for IP technology. MPLS Networks relay on sending data packets over Label Switched Paths (LSPs). The system of LSPs is similar to the concept of Virtual Path (VP) system applied in Asynchronous Transfer Mode (ATM) networks. In Synchronous Digital Hierarchy (SDH) networks the cross-connect for Virtual Containers (VCs) has to be set up, using as few resources and as short paths as possible. In Wavelength Division Multiplexing (WDM) networks or MultiProtocol Lambda Switching (MPλS) wavelength channels should be set up.

For these and other future types of networks an obvious model, the graphs are widely used; and the common representation of virtual paths applied by the above technologies is the graph theoretical term *path*. This model has the advantage that several earlier graph theoretical methods can be applied in telecommunication networks. The method of paths selection (Label Switched Paths or Virtual Paths or λ paths) is one of the interesting questions in traffic engineering. Paths can be either

static when the connection is permanently alive or *dynamic* if it is built up on-demand. Static paths are efficient in case of constant bit-rate while dynamic paths in case of bursty traffic.

The task of the *network manager* is to set up connections, keep track of them, and taking them down. The network management is either centralized in which case it is realized by a *network management center* or some functions are distributed. In the latter case every network element has a module that manages this element. These modules communicate with the network management center. To each network element to be managed belongs a *management information base* that contains the variables representing the information that are monitored and controlled by the network element manager, e.g., the type of the fiber, the maximal bit-rate and the type of protection.

Algorithms of graph theory are used in several areas of network design and management. In case of establishment of a point-to-point connection shortest path or minimal cost flow is searched depending on whether the connection can split. There exist more sophisticated methods to protect networks against failures, which determine two or more edge- or vertex-disjoint paths. A spanning-tree of minimum cost is required in case of establishing a point-multipoint connection and in most cases access networks are trees as well. The wavelength assignment problem in Wavelength Division Multiplexing (WDM) networks and the frequency assignment problem in mobile networks can be handled by graph coloring algorithms.

We use smaller font size to refer to the applications of the topics in network design.

1.10.2. Fundamentals

A **graph** is an ordered pair $G=(V,E)$ where V is a nonempty set and E is a set of pairs composed of the elements of V. The elements of V are called **vertices** and the elements of E **edges**. Two edges of a graph are **adjacent** if they have a common vertex and two vertices are adjacent if they are connected by an edge. The number of incident edges of a vertex is the **degree** of the vertex.

A sequence $(v_0, e_1, v_1, e_2, v_2, \ldots, v_{k-1}, e_k, v_k)$ is a **path** if e_i is an edge connecting v_{i-1} and v_i and all vertices are different. If the vertices are all different

except v_0=v_k then it is a **circuit** in the graph. If there exists a path between each pair of vertices then the graph is **connected**.

Networks are usually modeled by **directed graphs** (e.g., if the link capacity from node *A* to node *B* differs from the capacity from *B* to *A*). The edges of directed graphs are ordered pairs of the form (v_1, v_2) instead of unordered pairs of the form {v_1, v_2}. The vertex v_1 is the **tail** while v_2 is the **head** of such an edge (v_1, v_2)

The definition of **directed path** and **directed circuit** are analogous to those of path and circuit. A directed graph is **strongly connected** if there is a directed path from every vertex to every other vertex.

In many applications numbers are associated with the edges or vertices of the graph. These numbers represent for example the physical length of, or delay over a link, or the cost of using the link, or the delay or cost of the equipment. In this case the graph is called **edge-weighted** or **vertex-weighted** graph.

A graph is sometimes stored as a matrix. The **adjacency matrix** - indexed by two vertices of the graph - contains ones (or in case of weighted graphs the weight of the edge) if the two vertices are connected by an edge and zeroes otherwise. In case of directed graphs the direction of the edge is given by the sign of the numbers. In the **incidence matrix** - indexed by a vertex and an edge of the graph - the non-zero elements indicate that the edge is incident to the vertex. The directions of the edges are given by signs again.

1.10.3. Trees

An **acyclic** graph is one that contains no cycles. A connected and acyclic graph is called a **tree**. A tree that consists of *n* vertices has *n*-1 edges. In a tree any two vertices are connected by a unique path.

Hence the routing is trivial in networks of tree structure.

A graph *F* is a **spanning tree** of the graph *G* if *F* is a tree and its vertex set is identical to the vertex set of *G* and the edges of *F* are all edges of *G* as well. The **weight** of a tree in a graph is the sum of the weights associated with the edges of the tree. Every connected graph contains a spanning-tree; our aim is to find the one with minimum cost. The easiest way of finding it is to use the greedy algorithm that works

as follows. First, choose an edge of minimum cost. Suppose that some of the edges are already chosen. Then choose an edge of minimum cost that does not form a circuit together with some of the already chosen edges. If there is no such edge then the algorithm stops, otherwise the above procedure is repeated.

A spanning-tree of minimum cost is required in case of establishing a point-multipoint connection and in most cases access networks are trees as well. In such a case two additional constraints may occur: first, the degree of the vertices in the tree may not exceed a certain number because of hardware limits; second, the path-length can be limited due to delay and other signal damaging factors.

In many cases not all points of the networks are to be connected, only some distinguished points are to be included in the tree. Then a different type of tree is needed.

Designate some points of the graph, which are called **terminals**. The tree that contains all the terminals is called a **Steiner-tree**; the one of minimum weight can be of smaller weight than the one obtained by the greedy algorithm. The tree may contain vertices that are not terminals; they are called **Steiner-vertices**.

1.10.4. Paths

The connection between two points of the networks can be of two types according to the technology: Either the connection should be realized along a single path (described in this section) or the connection can split (Figure 1.11.1). The latter case is discussed in the next section.

In most cases one must find a minimum weight (shortest) path, i.e., the sum of the weights of edges are to be minimized along the path. The best known shortest path algorithm is due to Dijkstra, which is applicable both in undirected and in directed graphs and yields the distance between a given point and all the other points, if the weights are positive numbers.

Negative weights are not permitted in Dijkstra's algorithm but the algorithm of Ford allows negative weights as well, but only for directed graphs. In this case directed circuits with negative total weights are not permitted. The algorithm of Floyd determines the distance in such a graph from any vertex to any vertex.

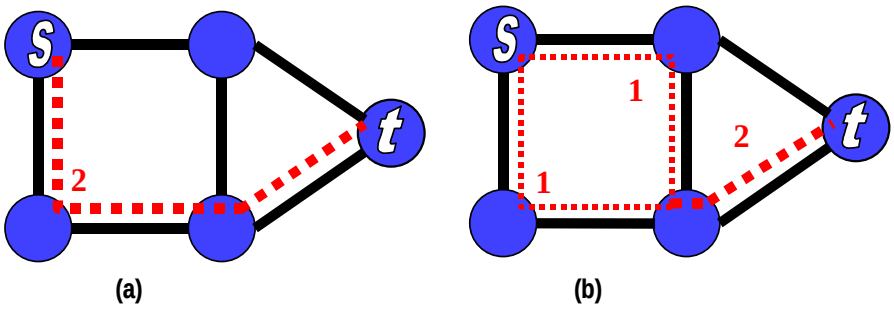


Figure 1.10.1. The connection s-t (a) is realized along a single path (b) can be split

Two paths are **edge-disjoint** if they do not have any common edge and are **vertex-disjoint** if they do not have any common vertex (except the end-vertices). Obviously, vertex-disjoint paths are edge-disjoint as well.

In survivable networks two or more paths are determined for the connections, which are edge- or vertex-disjoint depending on the type of protection.

There is a modification of Dijkstra's algorithm (also known as Suurballe's algorithm) that finds two or more edge- or vertex-disjoint paths of minimum total weight in an undirected or directed graph with positive weights, if they exist.

A ***k*-connected** graph has at least $k+1$ vertices and after deleting any set of less than k vertices the graph remains connected. In case of a ***k*-edge-connected** graph after deleting any set of less than k edges the graph remains connected. The graph is k-connected if and only if it has at least $k+1$ vertices and any two distinct vertices are connected by k vertex-disjoint paths; similarly, it is k-edge-connected if and only if any two distinct vertices are connected by k edge-disjoint paths.

In a directed graph the highest possible number of pairwise edge-disjoint directed paths equals the minimum number of edges intersecting all directed *s-t* paths. Similar statement holds for the highest possible number of pairwise vertex-disjoint directed paths if the graph does not contain an *(s,t)* edge.

1.10.5. Flows and Cuts

Let G be a directed graph. Associate with each edge e a non-negative number, **capacity,** denoted by $c(e)$. Designate two vertices s and t in G, called **source** and **sink**, respectively.

The capacity of an edge usually corresponds to the highest bit-rate (transmission capacity) of a link. Suppose that an amount of data (m_f) is to be transmitted from *s* to *t.*

Let *f(e)* be the amount of data flowing over edge *e*. The function *f* is **feasible** if for each edge $f(e) \leq c(e)$ and for each vertex the flow conservation constraint holds, i.e., the sum of *f* values on the out-going edges equals the sum on the in-coming edges, except for the vertices *s* and *t* where the sum of the values of the out-going and the in-coming edges should be m_f, respectively. In this case the function *f* is called **flow** and m_f is called the **value** of the flow. An edge is called **saturated** if $f(e)=c(e)$ and **unsaturated** if $f(e)<c(e)$.

The maximum value flow can be determined by the following algorithm. Suppose that there is a directed path from *s* to *t* so that for every edge on the path the flow value is smaller than the capacity, in other words, every edge of the path is unsaturated. Along this path the flow value of every edge can be increased until at least one edge will be saturated. The value of the flow can also be increased if the flow is decreased on an oppositely oriented directed edge. Such a path is an **augmenting path.** A theorem states that the value of a flow is maximal if and only if there is no augmenting path from *s* to *t.*

Divide the vertex set *V* of the graph into two subsets: let one set be *X* and the other one *V—X*. Set *X* is assumed to include vertex *s* and set *V—X* is assumed to include vertex *t.* An **(s,t)-cut** of a flow is the set of edges having one end-vertex in *X* and the other in *V—X.* The value of the cut is the sum of edge capacities on the edges pointing from a vertex in *X* to a vertex in *V—X*, i.e., only the "forward-edges" influence the value of the flow. The theorem of Ford and Fulkerson says that the value of the maximum flow equals the value of the minimum cut. The above-described method of augmenting paths finds both the maximum flow and the minimum cut if always the shortest possible augmenting path is chosen. If the capacities are integers, the maximum flow will be integer and it can also be realized so that each edge will have integer flow value.

In some applications numbers are associated not only with the edges but with the vertices as well, bounding the flow on the vertices from above. This problem can be reduced to the previous one in the following way. Substitute each vertex *v* of capacity *c(v)* with two vertices *v′* and *v″*. The new head of all edges pointing to *v*

should be *v′* and the new tail of all edges leaving *v* should be *v″*. Moreover, add a new edge of capacity *c(v)* with tail *v′* and head *v″* (Figure 1.11.2). The capacity of the new edge *(v′, v″)* expresses the capacity of vertex *v*.

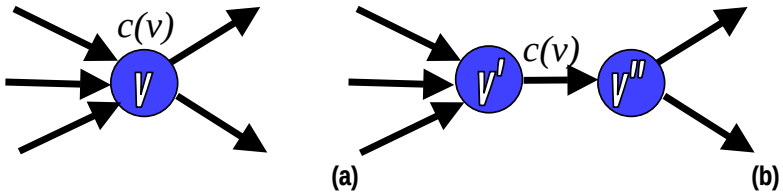


Figure 1.10.2. "Vertex splitting": Substitute (a) a vertex of capacity c(v) with (b) an edge of capacity c(v).

If undirected edges are permitted then they can be substituted by two oppositely oriented directed edges, that means two directed edges *(u,v)* and *(v,u)* of capacity *c* in place of a directed edge *{u,v}* of capacity *c*.

Up to this point, we assumed one type of data flow (i.e., one commodity), which is called **single commodity flow**. However, in telecommunication networks several connections are established which have to be distinguished (in general by their source and sink). These flow problems cannot be solved one by one because sharing of the common edge capacities binds the different commodities together, i.e., the sum of the flow values on each edge should not be greater than the capacity of the given edge. This is the so-called **multi-commodity flow** that is often only approximately solvable.

1.10.6. Graph Coloring

A graph *G* is ***k*-colorable** if every vertex of *G* can be colored such that the colors of adjacent vertices are different. The **chromatic number**, $\chi(G)$, of *G* is *k* if *G* is *k*-colorable but not *(k*-1)-colorable. The set of vertices with the same color is called a **color-class** of a coloring.

A graph is **complete** if there is an edge between every pair of vertices. A **clique** is a complete subgraph of *G*. **Clique-number**, $\omega(G)$, is the number of vertices in the maximum clique of *G*. Obviously, if there is a clique in the graph then any two vertices of it are of different colors, i.e., the chromatic number is not smaller than the clique-number: $\chi(G) \geq \omega(G)$. Let Δ(G) denote the highest degree in the graph. Let us color the vertices of the graph in arbitrary order. A vertex that is to be colored can

have at most $\Delta(G)$ previously colored vertices; accordingly, the $\Delta(G)+1^{st}$ color can be used for this vertex. That means the chromatic number is at most by one greater than the highest degree: $\chi(G) \leq \Delta(G)+1$. Hereby both lower and upper bounds are given for the chromatic number. However, these bounds are often far away from the real value.

The graph is **planar** if it can be drawn on the plane so that its vertices are distinct points of the plane and its edges are continuous lines between the end-vertices without common internal points. Every planar graph is 4-colorable: $\chi(G) \leq 4$.

A graph G is ***k*-edge-colorable** if every edge of G can be colored so that the colors of adjacent edges are different. The **edge-chromatic number**, $\chi_e(G)$, of G is k if G is k-edge-colorable but not $(k-1)$-colorable. The edge-chromatic number cannot be smaller than the highest degree since the edges that are incident to a single vertex are to be colored by different colors. However, for simple graphs the edge-chromatic number can exceed the highest degree by at most one: $\Delta(G) \leq \chi_e(G) \leq \Delta(G)+1$.

Graph coloring is useful in designing Wavelength Division Multiplexing (WDM) networks where wavelengths are to be assigned to each path. Paths using the same link cannot use the same wavelength. Consequently, we can construct a graph with vertices modeling the paths of the WDM network and edges modeling that two paths have common edges. The chromatic number of the graph gives the minimum number of required wavelengths and a particular coloring gives a proper wavelength assignment.

Another application is the design of mobile networks where certain stations (e.g., neighbors) may not use the same frequency. Consequently, a graph can be constructed with vertices modeling the stations and two vertices are connected if the corresponding stations may not use the same frequency. The chromatic number of the graph gives the minimum number of required frequencies and a particular coloring gives a proper frequency assignment.

1.10.7. Complexity of Algorithms

Most of the described algorithms (minimum cost spanning-tree, shortest path, maximum flow or minimum cut) are useful also in practice: the number of steps is in

the worst case upper bounded by a (generally low degree) polynomial of the input size. Thus, an implementation yields optimal solution in reasonable time even in case of a graph with thousands of vertices. It can be decided similarly in polynomial time whether a graph is connected or 2-connected; whether a directed graph is strongly connected; whether a graph is a planar graph.

On the other hand, there are many problems for which there is no known exact, polynomial-time algorithm. Among the aforementioned problems, the following ones have this property: finding a minimum-cost Steiner-tree, the longest path or the maximum cut; the general solution of the minimal cost multi-commodity flow, or the determination of the clique-number, the chromatic number or the edge-chromatic number. These problems belong to the class of **NP-hard** [1.10.5] problems: if somebody would be able to find a polynomial time algorithm for any of them then by calling that algorithm as a subroutine it would provide a polynomial-time algorithm for every other NP-hard problem. The running time of an NP-hard problem can be very long - depending on the size of the network and the type of the problem.

Solutions of this type of problems, even those which will be discovered in the future, will most probably have at least one of the following properties:

1. either the number of steps will not be a polynomial of the size of the input,
2. or they will not find the optimal solution, just an approximate one,
3. or a theoretically different model of calculation is needed (e.g., an algorithm requiring the generation of random numbers).

However, it is to emphasize that

a. if the size of the input is not too large then this distinction is not critical;

b. an NP-hard problem may have special cases that can be solved in polynomial time (e.g., the determination of the edge-chromatic number in a bipartite graph or finding the longest directed path in graphs without directed circuits);

c. there are problems for which no polynomial-time algorithm is known but - to our best knowledge - they do not belong to the class of NP-hard problems either (e.g., deciding whether two graphs are isomorphic).

1.10.8. Examples

Consider the network example in Figure 1.11.3. This network is modeled by the graph depicted in Figure 1.11.4.(a). In order to establish a connection a shortest

path is searched between two routers (Figure 1.11.4.(b)). This traffic can be protected in two ways: by *path protection* (Figure 1.11.4.(c)) or by *link protection* (Figure 1.11.4.(d)). In case of path protection each connection is protected one by one, i. e., the traffic is re-routed between source and destination avoiding the failed link. This approach saves link capacity, however the detection of failures can take a long time. A special case of path protection is when working and protection paths are disjoint. In this case there is no need to determine the exact place of the failure, accordingly the traffic can be immediately restored. In case of link protection just the traffic of the failed link is to be rerouted.

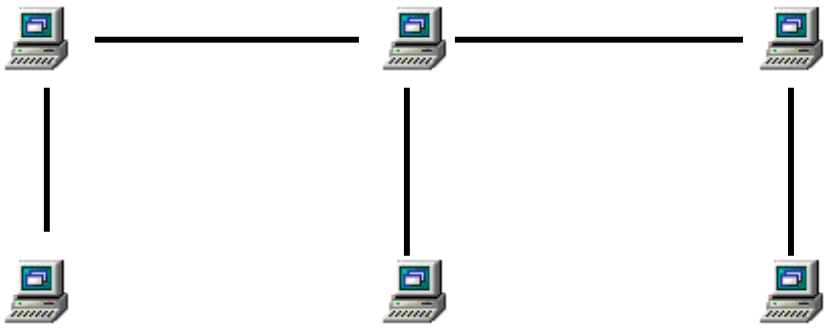


Figure 1.10.3. Telecommunication network of six nodes

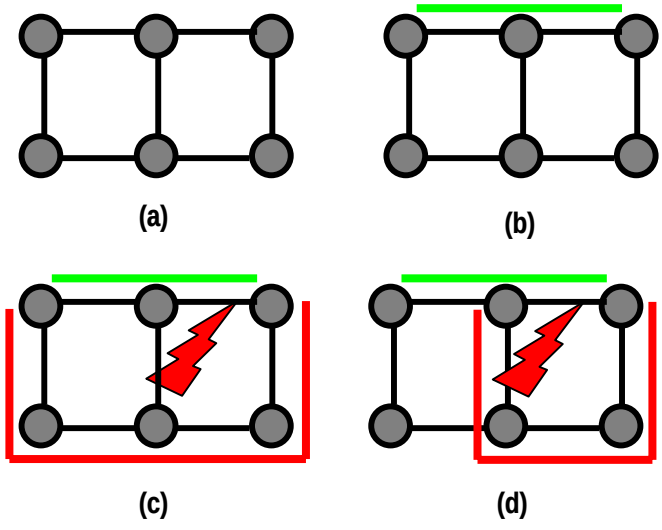


Figure 1.10.4. (a) Graph model of the network on Figure 1.10.4. (b) establishment of a connection, (c) path protection, (d) link protection

Figure 1.11.5. shows an example of a built-up Steiner-tree where terminals are marked with a circle. This Steiner-tree contains one Steiner-vertex; i. e., one additional client is needed to reach all terminals.

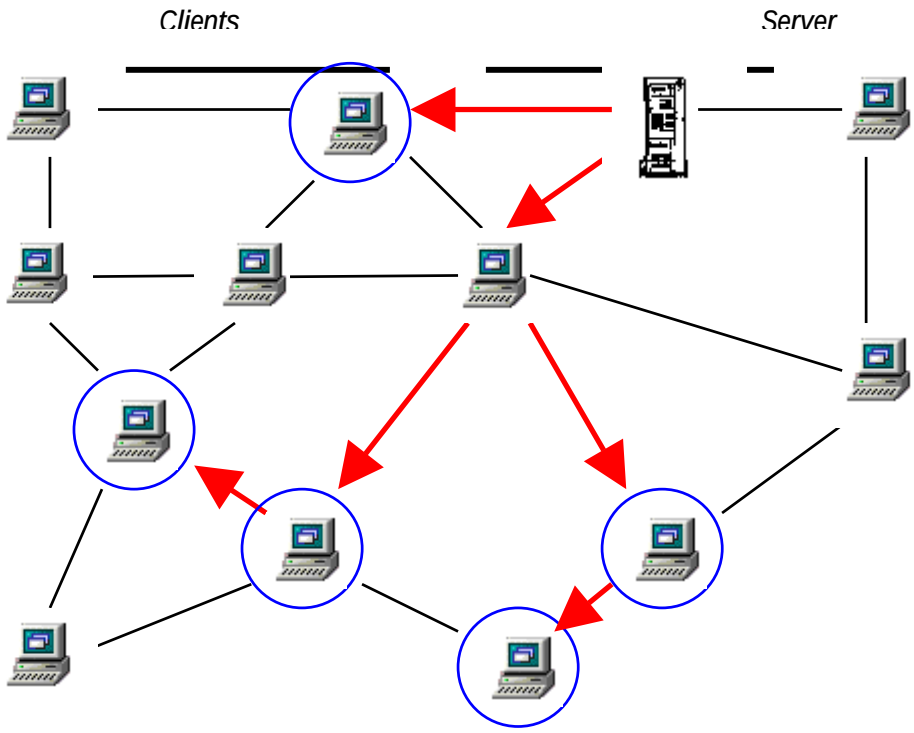


Figure 1.10.5. Steiner tree example

References

A good introduction to graph theory can be found in the books [1.10.1] or [1.10.2] and in the problem collection [1.10.3], while [1.10.4] is recommended for introduction to application-oriented students. Algorithms in graph theory are deeply covered in [1.10.5]. The handbook [1.10.6] is a must for active researchers in the field.

[1.10.1] J. A. Bondy, U. S. R. Murty, "Graph Theory with Applications", The Macmillan Press Ltd, London, 1976

[1.10.2] R. Diestel, "Graph Theory", Springer, Berlin, 1997 and 2000

[1.10.3] L. Lovász, "Combinatorial Problems and Exercises", North-Holland, Amsterdam, 1979 and 1993

[1.10.4] J. Gross, J. Yellen, "Graph Theory and its Applications", CRC Press, Boca Raton, 1999

[1.10.5] R. E. Tarjan, "Data Structures and Network Algorithms", SIAM, Philadelphia, 1983

[1.10.6] R. L. Graham, M. Grötschel, L. Lovász, "Handbook of Combinatorics", Elsevier, Amsterdam and The MIT Press, Cambridge, MA, 1995

1.11. Layered modell of networks

Sándor Mazgon, author

István Bartolits dr., reviewer

Communication protocols developed as communications technologies grow and expand, became and should evolve to a more and more sophisticated form as a result of the prevalence of man-machine and machine-to-machine communication beyond the traditional man-to-man one.

Open systems, like intercommunicating computers and similar intelligent data terminals need for data communication among them such sets and suits of communication protocols, which can serve (or support) all of the requirements of any computer application and user. International standards making organizations (International Standardization Organization, **ISO**, and International Telecommunication Union, **ITU**) established a model for this purpose, which provids maximal felxibility while taking into account all the communication tasks to be executed among data terminal equipment (computers, dummy terminals, etc.). This is called the model of **OSI**, Open Systems Interconnection[1], the most general functional reference model of networks of telecommunicating entities.

To establish the OSI model a set of principles were determined, the most relevant are the followings:

a. Create the model as a set of independent layers, each of which is a different set of abstractions.

b. Do not create too many layers as to make the system engineering more difficult than necessary.

c. Create a layer boundary at a point where the number of interactions across the boundary are minimized. Create a boundary where it may be useful at some point in time to have the corresponding layer interface standardized. Select boundaries at a point which past experience has demonstrated to be successful.

d. Create separate layers to handle functions that are different in process or technology and collect similar functions into the same layer. Create a layer

[1] ISO 7498-1 | ITU-T X.200: see [1.11.1]

where there is a need for a different level of abstraction in the handling of data, for example morphology, syntax, semantics.

The layers of the OSI model, standardized according to the principles, may be summarised by the following: (see also Figure. 1.11.1.)

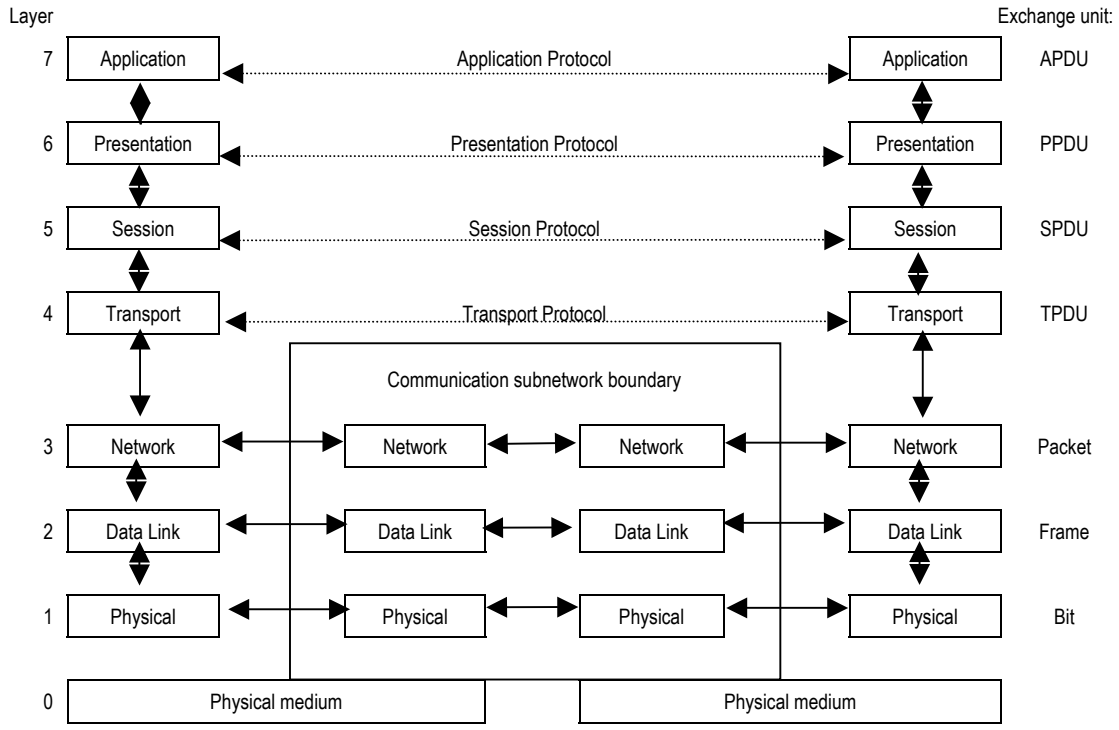


Figure 1.11.1. Communication architecture based on the OSI model

Physical layer: Provides electrical, mechanical, functional, and procedural characteristics to activate, maintain, and deactivate physical links that carry the bit stream transparently, independently from any data structure a it only recognises individual bits, not characters or multicharacter frames.

Data link layer: Provides functional and procedural means to transfer data between network entities, detect and (possibly) correct transmission errors; provides for activation, maintenance, and deactivation of data link connections, grouping of bits into characters and message frames, character and frame synchronisation, error control, media access control, and flow control (examples include HDLC and Ethernet). The data link layer masks error characteristics and noise patterns of the physical data communications circuits from the network layer.

Network layer: Provides transfer of packets (inserted in the frames of the second layer -- see Figure 1.11.2) between source and destination via the network connection. To execute this, layer entities have to got knowledge of the network structure and capability to select the optimal route is some sense, the procedure of which is called switching, relaying and routing. Provides independence from data transfer technology; masks peculiarities of data transfer medium from higher layers and provides switching and routing functions to establish, maintain, and terminate network layer connections and transfer data between users.

Transport layer: Provides equalization of any difference in quality of service provided by different network connections to the transport connection, provides transparent transfer of data between systems, relieving upper layers from concern with providing reliable and cost effective data transfer; provides end-to-end control and information interchange with quality of service needed by the application program; first true end-to-end layer.

Session layer: Provides scheduling of data transfer between end systems, provides mechanisms for organising and structuring dialogues between application processes; mechanisms allow for two-way simultaneous or two-way alternate operation, establishment of major and minor synchronisation points, and techniques for structuring data exchanges.

Presentation layer: Provides presentation of information for communication and management of data transferred in a unified manner, provides independence to application processes from differences in data representation, that is, in syntax; syntax selection and conversion provided by allowing the user to select a "presentation context" with conversion between alternative contexts.

Application layer: The highest layer in the OSI model provides access for the application processes to the OSI environment. This layer is concerned with the requirements of the applications. All application processes use the service elements provided by the application layer. The elements include library routines which perform interprocess communication, provide common procedures for constructing application protocols and for accessing the services provided by servers which reside on the network.

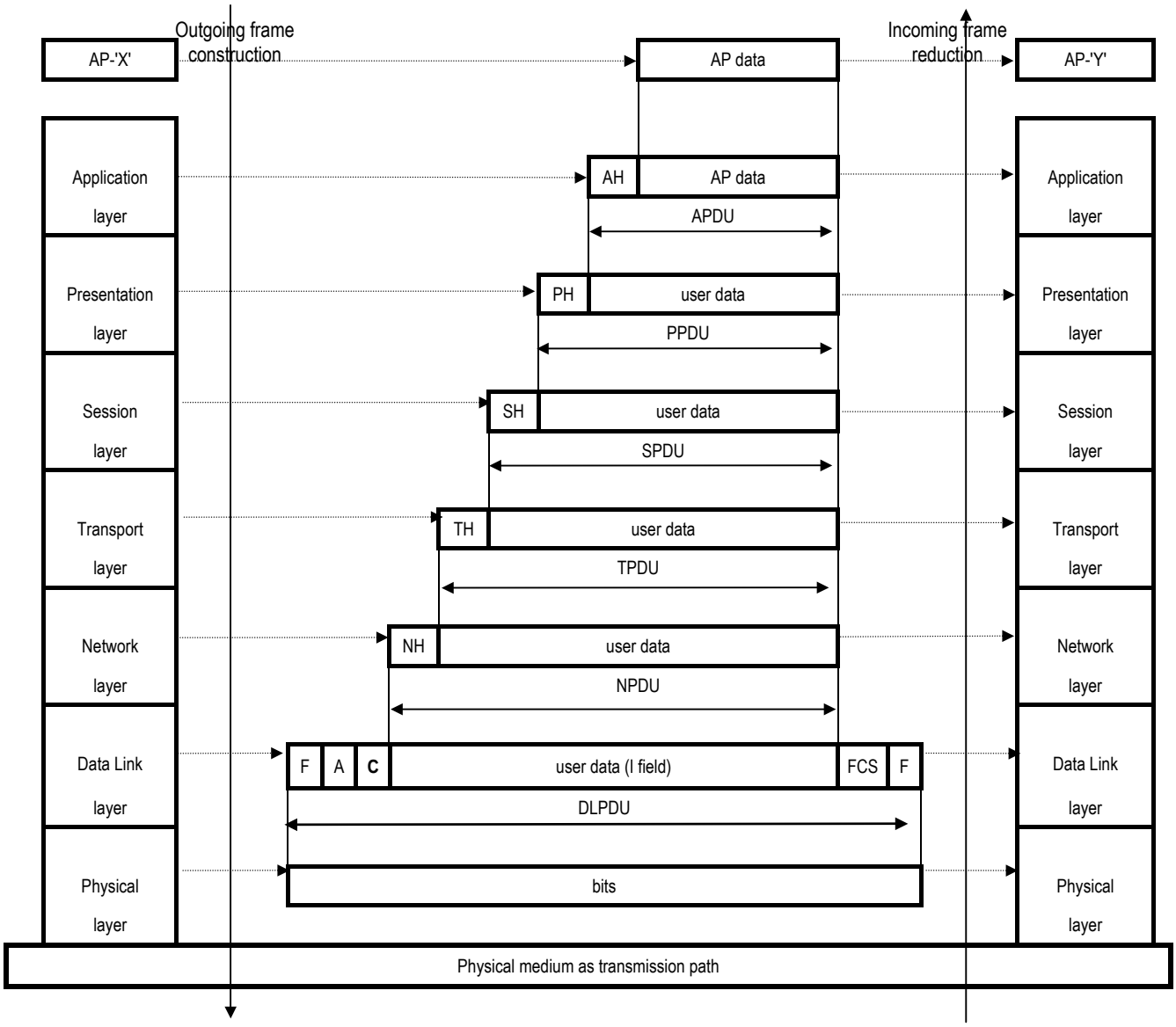


Figure 1.11.2. Packeting layer-by-layer in the OSI model

Legend:

- AP application program ('X' at one side and 'Y' at the other side)
- AH application header PH presentation header
- SH session header TH transport header
- NH network header FCS frame check sequence
- APDU application protocol data unit PPDU presentation protocol data unit
- SPDU session protocol data unit TPDU transport protocol data unit
- NPDU network protocol data unit DLPDU data link protocol data unit
- F flag A address
- C control I field information field

Note: bits are the physical protocol data units (PhPDUs) in the abstract physical layer and are carried by signals in the data circuits established in the physical medium.

In the model all the layer functions are executed by functional layer entities which communicate with other entities, with the peers in the same layer of other open systems by the use of protocol data units (**PDUs**) of relevant layer protocols.

Every communications between open systems are carried by data circuits routed in the physical medium in the form of data signals exchaged by the end systems. Communication is always originated by an application process represented in the application layer by an application entity, and all the layers are taking part in it. The application entity sends to its peer application protocol data units (**APDU**) consisting of application data (**AP data**) and accompanying protocol control information, the application header (**AH**) as represented in the **Figure** 1.11.2. Each layer appends to each **PDU**, it received from the next layer above, its own protocol control information to form its **PDU** and hands over to the next layer below. In that way the **APDU** is user data for the presentation layer, and with an appended protocol control information PH (presentation header) forms the presentation protocol data unit (**PPDU**). In the same manner from the session layer **SPDU** = user data (**PPDU) + SH,** from the transport layer **TPDU** = user data (**SPDU) + TH,** from the network layer **NPDU** = user data (**TPDU) + NH,** as exhibited in the **Figure** 1.11.2., where the horizontal arrows represent peer-to-peer protocols, and the vertical arrows represent layer interfaces. These protocols and interfaces are standardized in the OSI system (see X.200 series of ITU-T Recommendations).

OSI realisations exist in many variants, most are in the lower three layers. A very complex task would be the comparison and association among them. In most cases the networking (the lower) layers are analysed by the concepts of the OSI reference model, as a common language of the different systems.

References:

[1.11.1] ITU-T X.200(1994) | ISO/IEC 7498-1:1994 Information processing systems. Open systems interconnection. Basic reference model. Part 1: The basic model.

[1.11.2] ITU-T X.800(1991) | ISO 7498-2:1989. Information processing systems. Open systems interconnection. Basic reference model. Part 2: Security architecture.

[1.11.3] ITU-T X.650(1996) | ISO 7498-3:1997 Information processing systems. Open systems interconnection. Basic reference model. Part 3: Naming and addressing;

[1.11.4] ITU-T X.700(1992) | ISO 7498-4:1989 Information processing systems. Open systems interconnection. Basic reference model. Part 4: Management framework

Abbreviations (see also Figure 1.11.1 and 1.11.2.)

ITU, ISO, IEC, MSZ, idt, OSI, HDLC, AP, AH, PDU, APDU, PH, PPDU, SH, SPDU, TH, TPDU, NH, NPDU, F, A, C, FCS, DLPDU, PhPDU

1.12. Telecommunication economics

György Lajtha dr., author

Lászlóné Konkoly, reviewer

In the first century of the telecom history (1880 – 1980) telephony and telegraphy were the only public services in most European countries. In the last decade it changes essentially: new teleservices, entertainment and business applications has been developed. This difference has had asignificant impact on the economic optimization used in the planning process, and on the concept of dimensioning and maintaining a network. In the last twenty years there has been some other crucial changes too, just to mention the following ones:

- The ratio of man power prices became much higher comparing to electronic components. The land and housing costs are growing relating to the price on one channel or transmission path;
- The tarifs became relatively higher comparing to the price of the telecom equipment, as it was some decade earlier.
- A further influencing factor is that telecom was a state monopoly earlier but now it is a specific field of competition.

  All these changes are supporting our task to give on overview on this topic.

1.12.1. Classical Network Planning Procedure

In the world wide used procedure the first step is the construction of traffic matrices for example for 5, 10 and even for 20 years perspectives. This is the basis of finding the proper routing strategy. The result will be the network structure fitting to the traffic throughput requirements, it can be a star, a multiplestar, a bus, a tree, a ring or a double ring topology. (See Figure 4.1.1.) The choice depends also from the cultural, administrative, educational and health-care structure of the region. It defines also the places of traffic concentration points, the alternative routes and the availability (defined as the uptime related to the whole calendry time) strategy. When the traffic (logical) structure is defined, then the next task is the planning of the realization.

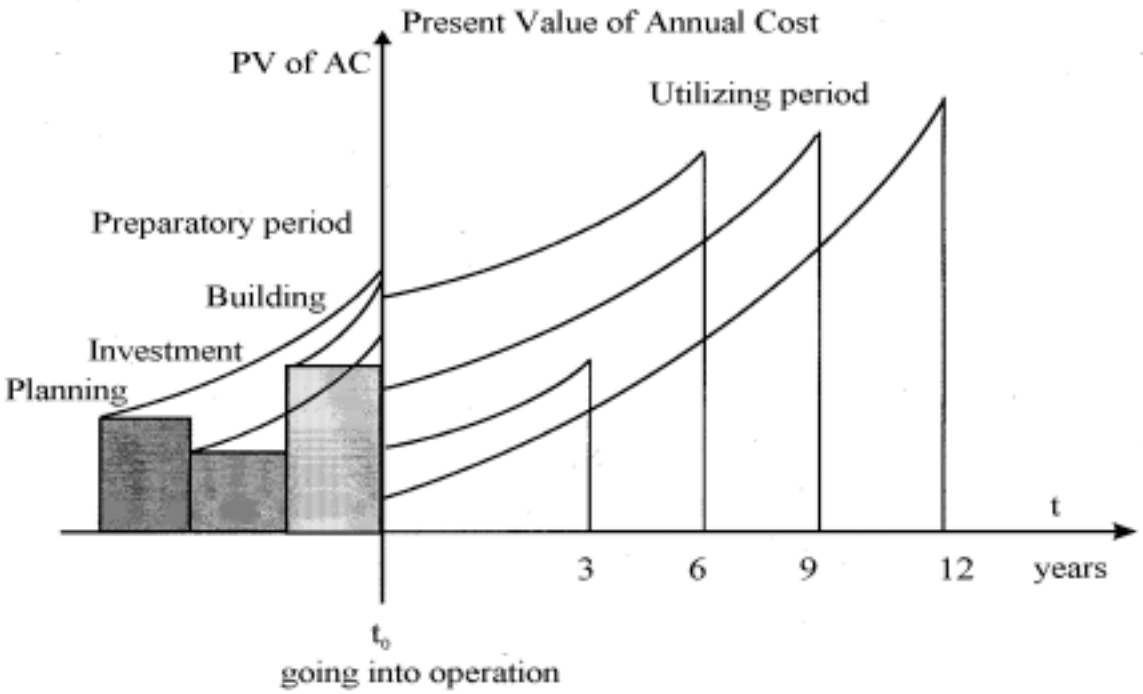


Figure 1.12.1. PV of AC curves

Be careful! The cost of planning and investments
before the beginning of operation is negative

The realization must take in to account the existing network (infrastructure), the obtainable equipments and the new plan must be harmonized with them. The economic optimum can be found with using the PV of AC method, which means that every income and expenditure is to be transferred on the time scale to the “going in operation” point which is the origo of the time-cost diagram (to). To the development-, planning-, building-, and mounting-costs must be associated negativ time so the PV of AC is higher than the real expenditure.

The PV of AC (Preset Value of Annual Cost), when we invest n_i capacity at time t_i is as follows:

$$K_i = \frac{C_0 + C_{n_i} n_i}{(1+r)^{t_i}} + \left[S_0 + n_i S_{n_i}\right] \frac{(1+r)^{T-t_i} - 1}{(1+r)^T \cdot r} - \frac{M}{(1+r)^T} \qquad \text{....1.12.1.}$$

$$K = \sum_i K_i \quad \text{and} \quad \frac{1}{(1+r)^{t_i}} \rightarrow \begin{array}{ll} > 1 & \text{if } t_i < 0 \\ = 1 & \text{if } t_i = 0 \\ < 1 & \text{if } t_i > 0 \end{array}$$

Here M is the residual valoe which can be realized at the and of the planning period T; K_i denotes the PV of the investment done in $t=t_i$, C_o is the capacity independent part and C_n is capacity proportional part of the investment, S_0 and S_n are the costs of the similar parts of the maintenance. T the nominal "life time" of the system can be forexample 5, 10 or even 20 years. r is the telecom interest rate which is higher than the bank interest, due to the higher risk, the short obsoletion time, and the not fully exploited systems. S_o $r_{telecom}$ can be 10-25 % when r_{bank}=4 %. In general an investment is realized in more steps, so the total PvofAC is

$$K=\sum_i K_i \qquad \text{....1.12.2.}$$

1.12.2. Business oriented optimization

The company's owners want to maximize dividends. The question is for what time period they wish to maximize their gain? In each year different dividends can be achieved, so *D(t)* is a function of time, and the present value of sum of dividends which can be expressed in quasi continuous case is:

$$D=\int_0^T D(t)\cdot\frac{1}{(1+\alpha)^t}dt \qquad \text{....1.12.3.}$$

Here D(t) is a function of incom.

The dividend *(D)* in a give year (t) is proportionate to the difference between income *(B)* and the expenses *(K),* and they can expressed in any monetary unit (MU)

$$D(t) = \beta[B-K] \qquad \text{....1.12.4.}$$

In any time the dividend is a given percent of the difference between the income and the costs. β is the function of the vecision of owner. It depends what amount is necessary for R+D, investments and replacement, of obsolate items of the network.

Where the value of depends on the owners' economic, technical and political decision and the company's strategy, but in general $\beta\leq 1$. The value of B depends from the usage and the tariff. Starting a service the usage is quite low. The income will approach the planned value only after 3-5 years. (Figure. 1.12.2.) The income is proportional with the traffic A, and the tariffs b, further on b is depending from the

value of service b (λ). Here λ covers the offered bandwidth, the guaranted quality

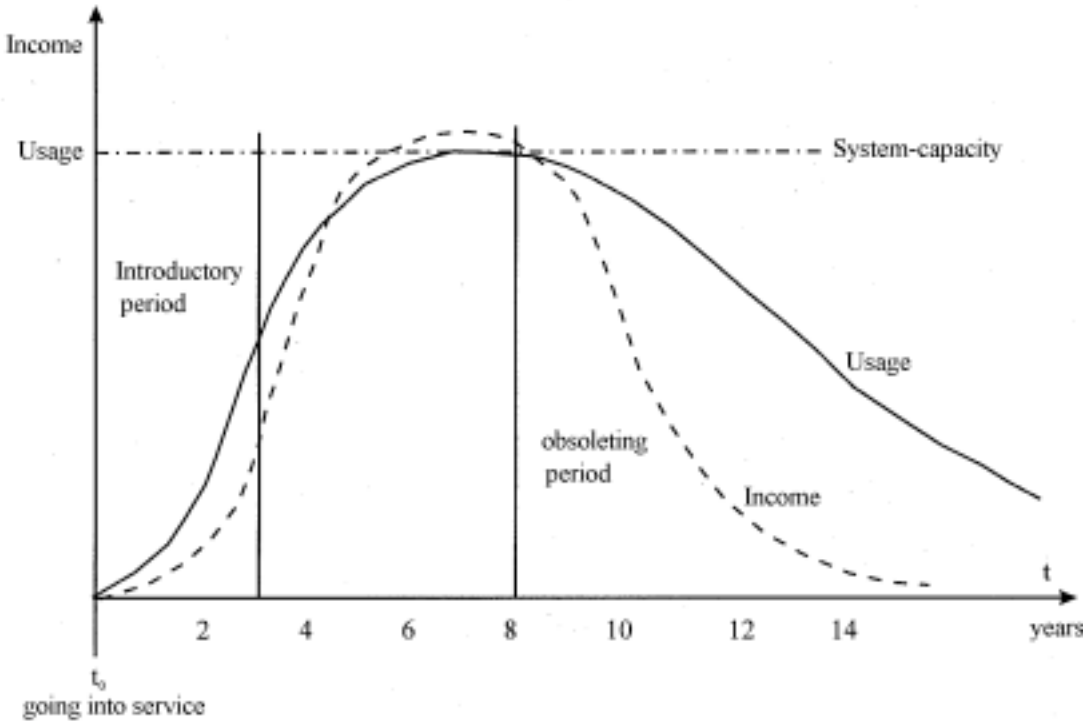


Figure 1.12.2.

and availability and the felxibility of the service provider. In most cases it is reasonable to decrease b (tariff), when the traffic reached its maximum.

1.12.3. The lifecycle of the telecommunication investments

The dimensioning of the network is influenced by two fundemental decisions of the owners

- The first one is the determination of time period in which they want maximize their gain or in an other explanation what should be the amount of divident or what percent should be β (1.12.4.)
- The second one is the determination the price of the usage (tariffs). This depends from the strategy business concept of the owners. The lower tariffs enhances the usage when the system became absolae.

Taking in account this factors it can be fixed that which should be the amount of network capacity in any part of the network. Matching to the definedquantity and quality requirements can be optimized the structure of the network, the place of nodes and the method of traffic routing [1.12.1., 3., and 7.]

Having the economic optimum we realize the network. The first step of the realization is the definition of the phisycal routes. This is influenced by the geographical circumstances, by the required availability, by the redundancies and spare capacities. Finally the maintenance system must be designed [1.12.1-7.].

The business orientation means that the below listed ten points must be taken in to account. These influence the amount of investment, the tariffs, the type of services.

- New technologies: packet switching, connectionless, democratic, -demand controlled routing,
- Different type of information: speech is not dominant, the data, picture, and the t required technology-incorporated telecommunications is enhancing
- Network organization: the hierarchical structure is replaced by the more versatile democratic one
- Transmission routes: the new media has quasi unlimited capacities
- Teleservices: the importance of the higher 4 layers (OSI hierarchy) became more important
- Competition: several service providers and operators are fighting for the users and for the networks
- Regulation: the style of it is changing, general not strict in details
- Investment components: software is determining the prices
- Shorter obsolation time: The need of high usage is necessary
- Revenue limit: competitors are making a threshold but the high interest rate requires higher prices.

Scrutinizing the changes and its effect we can learn that in the present time to transport bits is a simple task and not a good business. Telecom companies have to offer more sophisticated services satisfying the specific demands of the customers. So business goals must be redefined. Here the availability of the communication and the personal services have high priorities.

1.12.4. Competitive enviroment

Eonomic cousiderations are influenced not only by the users but also by the gross national income, by the strategy of the competitors and by the legal conditions. There was a quite general rule that maximum so about 5 % of the family's income to be spent for telecommunications was the target. Now it is also valid for the traditional services, but there are new services for which customers are ready to pay more.

The users' decision depends also on the competing services and service providers. Not only the price but also the quality, reliability and the speed of establishment of services make up this factor. Finally, framework conditions are dictated by the fixed environment that comprises legal and technical possibilities. (Figure 1.12.3.)

Beating the competition does not necessarily mean the maximum profit. It is only worth winning over users if the conditions are such that this does not mean losses for the company. Thus here the strategy needs to examine the social, economic and technical conditions in order to lead towards the goal set. Generally the economic goal is the integral of time-dependent profit over a given time period. Therefore, again and again we may repeat our decision process. In Figure. 1.12.4. we show what consequences will have a decision taken in the present. Later, at timepoints t_1, t_2 and t_3 decisions could be repeated and at the end of the total investigated period (T) we reach different results depending on the strategies of the other participants of the game and on the changing circumstances. The conclusion of Figure. 1.12.4. is that in every decision made today is the most important to prepare us to every possible further situations. The present step should not close the gate for any perspective future.

Competition is very useful because companies will be forced to become more creative and invest more money into research, development and production. Therefore most goverments support that there should be more than one player

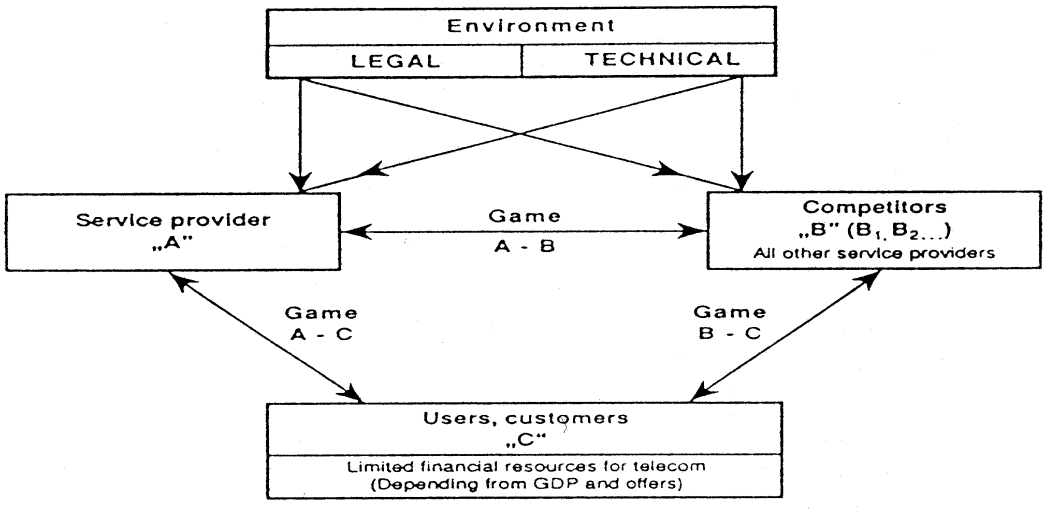


Figure 1.12.3. Players of the telecommuinication game

present on the telecom market. If it is realized and there are several service providers and networks operator then the situation can be described by a game, which supports the development of telecommunications and informatics of the conutry.

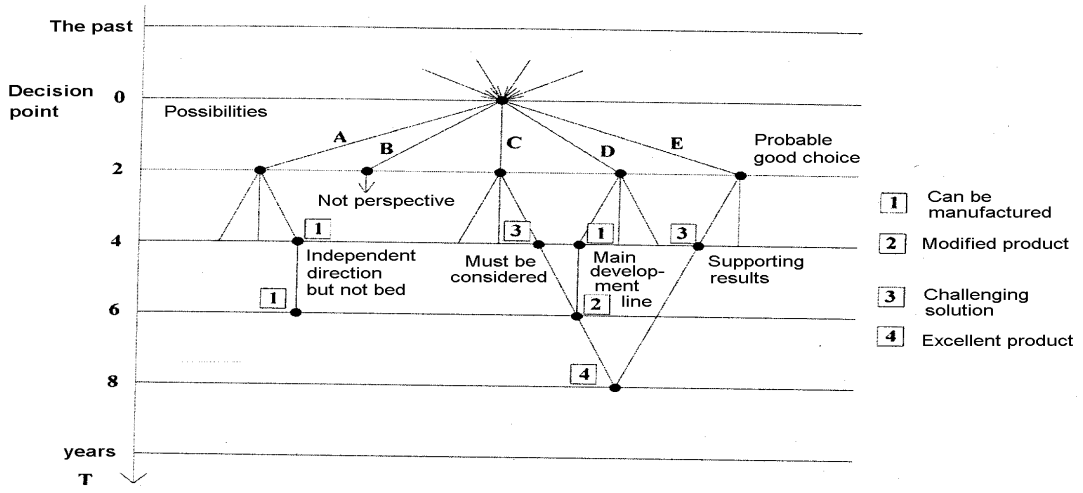


Figure 1.12.4. Consecutive steps of decisions

1.12.5. Game theory

The concept is that players want to maximize their profit, using the terminology of Game Theory: to maximize the pay-off function. If there are only 2 players we call it a duopoly model. In several cases it gives a result where both players found the best solution. In general there are more than two players. In several cases to simplify modeling and calculation we can, or have to eliminate the ones with minor importance. If more than two players have determining role in the game we can use models which are called quantity setting or quantity adjusting oligopoly models. N companies produce a homogeneous product and sell it for a unit price. The price depends on the total product output of the industry.

In a competitiv enviroment the companies participating in the game. have a comon aim: to enhance the total telecom expenditure of the population and companies. On the other side each competitor is fighting for more customer and more traffic on the given market. So there are some common purposes which means there is also necessary to cooperate to enhance the overall amount spending by the society on telecom services.

Game theory has been proved very useful in analysing industries, such as telecommunications, where there is only a limited number of companies being present in the market and there are a lot of externalities originated from the competitive environment. Game theory applications are able to take into account these effects of uncertainties and so can support the investment decision process of a company. It can be applied on different strategies where competitors are separated, or there is some information exchange or limited co-operation.

It is important to mention at the outset that game theory is not an analytical game of chance but is for solving strategic situations. We can analyze situations in which decision makers are not facing an objective environment but the decisions of other rational beings. A knowledge about who the other competitors are and what are their goals and abilities is obviously affecting the choice of our optimal strategy.

The players try to maximize their pay-of-function (P) which can be expressed with PV (PV of AC) of B-K (1.12.3., 1.12.4). There are some models for scrutinizing the output of the game.

The general characterisation of a game

- there are players who interact with each other
- there are rules specifying the permissible actions to be followed by the players
- there are strategies, which are decision variables dominant in the competition and which determine the outcome of the game (behaviour plans concerning production quantity, pricing policy, size of network, advertising of other decisions)
- payoff-function definition (which shows the outcome of the game, may be expressed e.g. by conventional PV of AC, PV of cash flow, extended (or adjusted) divident, value of the company, including option value)
- definition of the concept for the solution of the game.

Game theory focuses mainly on two solution concepts. These are the Nash equilibrium and the sub-game perfect equilibrium:

- A Nash equilibrium is an outcome of the game, in which no player wishes to change his strategy, because this unilateral change would cause for him a worse solution. This means that all the players take the best possible action given the actions of the others. This concepts is commonly used for normal form games. In this case players move simultaneously and only once.

- A sub-game perfect equilibrium is a refinement to the Nash concept. In this case a game has more than one stage or round. Such games are interpreted

in extensive form (given by a tree structure). The most common solution concept for extensive form games is the sub-game perfect equilibrium concept. (A sub-game is defined as a part of the game having all the characteristics of a game.) This method is preferable in the case of dynamic games, where more consecutive steps must be forseen. To which supports the solution [1.12.10.]

There is a practical limit in the game the solvent demand. It means that an important task is for each competitor to achive a high value of extra services what can be done by attractive service offer.

If every player has chosen his strategy, the game is already played and it leads to a well defined solution. The following paragraphs are demonstrating two quite common situation the practical use of Game Theory [1.12.8-14.].

1.12.6. Practical use of Game Theory (Example 1)

For the sake of simplicity suppose that the number of players is 2 (duopoly case). Both companies want to build an ATM network on their existing SDH systems based on permanent connections. On this interpretation of the Nash equilibrium each firm observes the quantity the other firm produces and calculates how much it should produce to maximise his profit. In this model we choose quantity of bandwidth as optimization parameter of the strategy. Let x be a measure of the ATM network capacity.

Suppose that the investment costs of the 2 companies are linear functions:

We suppose that the measure-independent parts are equal, but the measure dependent parts are different. This can be the consequence of the size economy, namely we suppose that the second firm. It proves that a higher amount of services can be realized only by lower unit price. Of building more extended ATM network. We denote the capacity by x_1, the second firm can it realize with less unit price due to the existing backround and the bigger capacity to be build. The capacity independent component is equal (32). The O&M costs are supposed to be involved (as a first approximation) ϑ in the investment costs. So we can calculate with the following equations:

$C_1(x)=0.08x+32$

$C_2(x)=0.04x+32$ ….1.12.5.

These consuming costs contain the costs of operation (energy consumption, training costs, wages, foreign conferences costs, software, licence fees, and so on), marketing costs (training and foreign trip costs, advertisement costs, market research costs, cost of polls, mailing costs and so on).

We suppose that the inverse demand function (unit price) has the following form: $p(x_{1,}x_2)=0.75-0.00005(x_1+x_2)$. It proves that a higher amount of services can be realized only by lower unit price has the capacity. We wish to maximise the profit of the companies through a given time period (T).

The solution of this optimisation problem can be shown in the ϑ Table 1.12.1., which also shows the possible strategy combinations together with the realised pay-off values. The upper row shows the strategies of player 2, and the left column shows that of the first. The cells of the table contain the pay-off function values for the companies 1 and 2 respectively. We can see that in case of the defined data the point (3200, 4400) is a Nash equilibrium, where the value of the pay-off function is of the two players 210, 666 respectively.

	2800	3600	4400	5200	6000
2000	298 706	218 850	138 930	58 946	-22 898
2400	370 650	274 778	178 842	82 842	-14 748
2800	426 594	314 706	202 754	90 738	-22 598
3200	466 538	338 634	**210 666**	82 634	-46 448
3600	490 482	346 562	202 578	58 530	-86 298
4000	498 426	338 498	178 490	18 426	-142 298

Table 1.12.1.

So, firm one who is capable to build smaller network has the strategy 3200, and the sompany who is willing to build bigger network has the strategy 4400 as an optimal choice.

There is a simple graphic method that can be used in duopoly cases. (Of course because of the dimension increase this graphic method can not be used easily when there are more than 2 firms.) The Nash equilibrium can be found by using the reaction curves. The intersection of the two curves gives the equilibrium point. The reaction curve can be derived from the derivatives of the pay-off functions. Fixing the first firm's output, we maximise the output of the second firm and vice versa. So, the reaction curve shows how a company behaves as a function of the

behaviour or the other. The curves can be sketched using the derivates of the pay-off functions.

To this example Figure 1.12.5. shows the reaction curves (they cross in E), and the points A,B, and C shows the points how the firms approach the equilibrium state. Starting from A, company 2 reduce its output in order to reach its reaction curve at B. Then firm 1 increases his output, moving to C and so on. They move towards E, which will be a stable point for both of them.

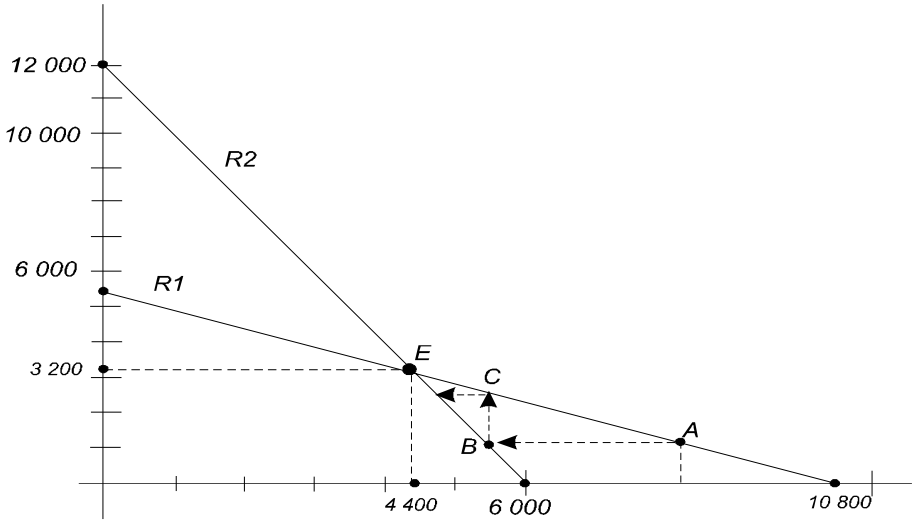


Figure 1.12.5. Finding the equilibrum with the use of reaction curves

A reaction curves of the firms are as follows:

R_1: $x_1=-0.5x_2+5400$

R_2: $x_2=-0.5x_1+6000$

1.12.7. Practical use of Game Theory (Example 2)

In telecommunication service provisioning there are several players, but we examine now a situation with two dominant players (duopoly). We suppose $\underline{x}$ nad $\underline{y}$ are offering services. The regulatory body is limiting the price of unit traffic (b). On the other hand for less than *m* it is not worthwhile to offer the service:

$m \leq b_{\underline{X}} \leq M$ and $m \leq b_{\underline{Y}} \leq M$

The usage (traffic, A) is decreasing with growing tariffs.

$$A\left(\frac{b_x+b_y}{2}\right)\rangle 0 \qquad \text{....1.12.6.}$$

The traffic distribution is a function of prices: **f ($b_{\underline{X}}$-$b_{\underline{Y}}$)**.

It is trivial that if $b_{\underline{X}}=b_{\underline{Y}}$ then f(0)=½.

According to eq 1.12.6. the income as a function of the unit tariff for <u>x</u> and <u>y</u> is:

$$K_x = b_x\, f(b_x-b_y)A\left[\frac{b_x+b_y}{2}\right] \qquad \text{....1.12.7.}$$

$$K_y = b_y\left[1-f(b_x-b_y)\right]A\left[\frac{b_x+b_y}{2}\right] \qquad \text{....1.12.8.}$$

We suppose now, that **m=1**, **M=4**, and further on the average traffic

$$A\left[\frac{b_x+b_y}{2}\right] = 20-(b_x+b_y) \qquad \text{....1.12.9.}$$

the distribution is

$$f(b_x-b_y)=\frac{1}{2}-\frac{b_x-b_y}{6} \qquad \text{....1.12.10.}$$

Now we calculate the income for <u>x</u> and <u>y</u> for the cases $b_{\underline{X}}$=1,2,3,4 and $b_{\underline{Y}}$=1,2,3,4.

The result is the following.

	b_y	1	2	3	4
b_x					
1		9 9	$11\frac{1}{3}$ $5\frac{2}{3}$	$13\frac{1}{3}$ 8	15 0
2		$5\frac{2}{3}$ $11\frac{1}{3}$	16 16	20 15	$23\frac{1}{3}$ $9\frac{1}{3}$
3		8 $13\frac{1}{3}$	15 20	21 21	26 $17\frac{1}{3}$
4		0 15	$9\frac{1}{3}$ $23\frac{1}{3}$	$17\frac{1}{3}$ 26	24 24

9, 16, 21 are equilibrium point. 24 would be the best but either *x* is or *y* can move to 26 but in that case the other compating party has a worse situation which he would like to change so it is an optimum but not equilibrium. The equilibrum can be modified also by quality, availability and the PR work. The choice of the three equilibrum paints depend on the user behaviour. It can be developped from 9 to 21. [1.12.8-14.]

1.12.8. Risk management

Every economic decision has some risk. This decision can be an investment, installation fo a new service, any kind of buying, selling, rental business, or financial transaction. The success depends on several influencing factors, which can not be precisely forecasted.

During the Course of preparation it is necessary to make some calculations which are based on the most probable situation. The difference between the present value of the income (B) and the costs (K), estimated for T=3….10 years period, can be considered as a result of the planned action (Ch. 1.12.1.):

$$\overline{E} = \overline{B} - \overline{K}$$

But these quantities are rather uncertain. The mean values can be affected by several factors which depend not only on the company preparing the action.

The decision can be supported by the probability density function of E, denoted by p(E) and sketched on 1.12.6.

To find a proper curve fitting to the circumstances it is a neccessary to collect the most probable risk factors

- success and strategies of the competitors
- the rules and limitations fo the anthority
- the behaviour, culture and economic situation of the users
- choice of the structure, technologie, availability of the the planned system, action on network
- deviation in the economic situation

A careful scrutiny of all the influencing variables the scale of Figure. 1.12.6. can be defined. Looking on the resulting figure we can the decide the following steps. If the probarbility of the action gives a negativ E value namely the action is not

profitable, to than it is not useful to continue the work in that direction, so we have to evaluate the equation:

$$P(0) = \int_{-\infty}^{0} p(E)dE$$

if P(0) is greater than a previousgivenly ε value, then the whole project must be rejected, or the management has to look for methods to reduce this probability.

In general the analysis and the reduction of risks has 4 important steps:

a. Collection and definition of the risk factors. It depends from the project itself, and also the legal and economic circumstances. Generally it is sketched in Figure 1.12.3. A possible list of $r_i$factors can be found in the introductory part of Ch.1.12.7.

b. Characterizing the risk factors can be done by their probability dansity functions and tiem dependence. In general we denote it by d (dependability) where

$d(r_i)=p(r_i,t)$

The density function of r_i is a similar curve to Figure. 1.12.6.

In some cases this function can be given in a deterministic form, but more often subjective estimation is necessary. Every subjective statement has some

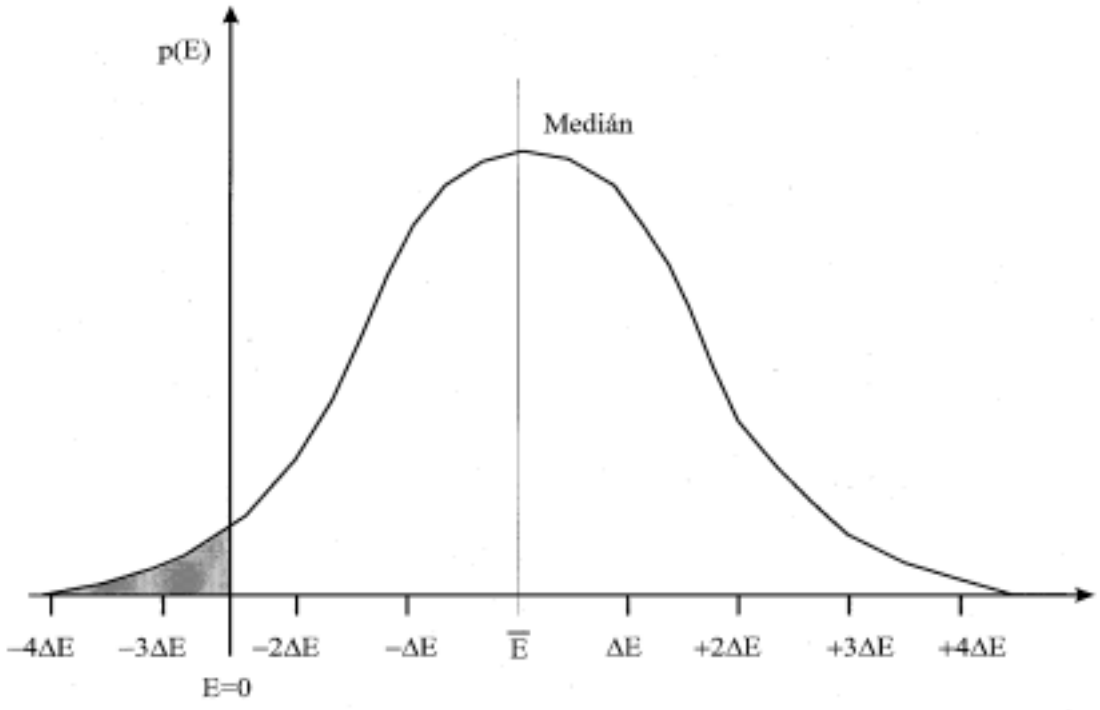


Figure 1.12.6.

incertainity. There exist appropriate methods to minimize this. Later in this point we discuss the problem.

c. Analysis of the impact of the risk factors on E. To find the relation between the changes of E and that of the r_i, we introduce the concept of sensitivity. Its definition is af fallows:

$$S_i = \frac{\partial E}{\partial r_i} \approx \frac{\Delta E}{\Delta r_i}$$

As a first step we try to collect all S_i values. For hardware, or any other systems with cansality in its evaluation, can be done using any objective law. By subjective opinion score (brainstorming). The r_i components for which is the largest, must be handled most carefully.

The second step is to construct the p(E) as a function of S_i and $d(r_i)$. a formal solution can be used. In some more sophisticated problems our aim is to find the analogy with the earlier mentioned process. In most cases p(E) can be drawn by computer aided methods (e.q. Monte Carlo simulation).

d. Reducing the negativ part of the density function and minimizing ε, so that ε<0,001E, is the last step of the process. Listing the greatest $d(r_i)$ values, the experts try to find methods to reduce them.

Better result can be achieved by scrutinizing S_i, by evaluating offers of further customer using different financial support, preparing the market. Using all there methods one com get in more efficient way smoller sensitivity. It is also useful to subdivide some of the risk factors, involving insurence companies or beguin the work (investment development etc.) in two alternativa direction and make the final decision only later.

Repetead use of the risk analysing process can enhance the probability of the technical and economic success of the proposed action. The consequence and the concardence test are useful in improving the *subjective methods*.

The consequence method, can be used by a team where several experts from all interested fields are participating. To enchance the reliability of the results of subjective test there are two methods. They are the following Consequence test wher each problem will be prepared as different kind of quistions. The most popular is the Quilford method, where on the same problem the experts are asked to perform the following 4 different evaluations. Let say there are five possibilities.

- From the five possibilities must be written in 10 pairs, and in every pare the answer should be a single preference statement in every 10 pairs.
- The 5 possibilities should be ranked in a single preferince list.
- At the third trial the first task is to delete the worst solution, it follows the declaration of the best one from the remeining 4. Giving some evaluation number to the other three it must be given 2 equivalent pairs.
- After these they have to position the 5 possibilities on a scale from 0…100.

If the four ranking methods give the same result, the consequence value of this expert is 1, if the result on every ranking is different, the consequence is 0. In the evaluation the opinion of the experts can be weighted by his own consequence value. Experts with 0 consequence value must be deleted from the group.

The *concordance* test is the second measure of the evaluation reliability shows that the ranking is the same by every experts (this is concordance=1) or every experts has different opinion (0). In general the number of the experts giving the same rank (N_S) divided by the total number invalued in the opinien score

$$Conc = \frac{N_s}{N}$$

If there are n values and m possibilities then n.m=N points are where the group gives a ranking value. Looking points we collect these points where the experts were giving the same ranking. It is N_S. It at least ⅔ of the points get the same number the concordance is acceptable if not it must be repeated with a completaly other team.

The course of risk management needs a lot of time and money. Therefore to apply this prodedure is useful only in such situation where the impact of risks is much higher (expressed in money) than the cost of the procedure.

1.13. Network and service quality (QoS=Quality of Service)

Gyuláné Nándorfi dr., author

Péter Kesselyák, reviewer

1.13.1. Quality trends (in the 21-st centuries)

Quality is a philosophical category. The essential determinant of objects. The charecteristic of an object, as a whole. The quality has already been one of the categories of Aristoteles. But, in the philosophy of Kant, the quality is a concept of higher level. It means, that the philosophy and the quality are in close connection.

Explanation and importance of the quality changed fundamentally in the last half century, depending on the change of market characteristic. In the era of monopolies the concept of quality has remained constant and its importance was ranked behind of quantitative development and price. In the liberalized market a change of paradigm has happened. On the marketplace of the products and services the concept of quality has become application-dependent and together with the price as a factor, determines the competition.

The main areas of international development, influencing the quality matters, are

- the new business models,
- the new technology and
- the new customer-oriented, quality-driven marketplace.

The new technology has an influence, first of all, on the network quality, while the customer-oriented quality rather is in connection with the quality of service.

Another important aspect, influencing the development of quality issues, is the integration of economical viewpoints into the accounting and financial management systems.

1.13.2. Concept of network- and service quality

One of the newest and most general definitions of quality, derived from ISO (ISO 9000:2000 DIS) was born in year 2000: the quality is "ability of a set of inherent characteristics of a product, system or process to fulfil requirements of customers and other interested parties" The product may be a traditional product or a service, too. The services are intangible, inseparable, perishable and variable.

The differences between the products and services lay in their quality interpretations, too. The quality of products is characterized, first of all, by dependability (reliability, maintainability, maintenance supportand availability), as well as efficiency. The quality of a network (as a product) is determined by the average quality of all connecting elements of the network, or by the quality of the worst link. The connection quality is determined as "the degree to which a connection reproduces the offered signal under stated conditions, as a sum of transmission performance, service retainability performance, propagation performance" (ITU Terms and Definitions Database "SANCHO") The Quality of Service is (ITU-T Rec. E. 800) "the collective effect of service performances which determine the degree of satisfaction of a user of the service". Especially, in "the telecommunications the Quality of Service is characterized by the combined aspects of service support performance, service operability performance, service integrity and the serveability performance".

The definition of QoS can be given either from the customer/user point of view, or from the service provider point of view. From the customer side, QoS is determined by the criteria being important for him. According to the service provider aspects, QoS may be characterized by parameters, being relevant to the end-to-end quality. QoS differs from network performance. QoS is the outcome of the user's experience, while the network performance is determined by the performances of network elements on-by-one, or by the performance of the network as a whole.. However, the network performance have an influence on the QoS, (it represents a part of it), through the service performances.

In the 1980-ths there was a general tendency, that the definition of quality and also that of the QoS, should be very general. But in the last decade came a new approch, proposing the application-dependent definition. A good example is given by

the definition of Eurescom P806 project, according to which the "QoS is defined as the degree of conformance of a service delivered to a user by provider with an agreement made between them."This QoS definition approaches the concept of conformance, what is not acceptable. Yet, this definition is in good coincidence with the final definition of ISO 9000:2000,according to which the quality is " the degree to which a set of inherent characteristic fulfils requirements". In definition of Eurescom appeared the concept of SLA (Service Level Agreement), namely the concept of the contract between the customer and provider. In the frame of SLA the provider assumes a guarantee for a defined level of quality (apart from the case of a service with non-guaranteed QoS).

1.13.3. Quality of Service models [1.13.2]

With help of a model, it is possible to analyze inherent relationships and the most important characteristics of the examined system or process. Models of QoS may be categorized from different aspects, for example according to their origin. The *Table 1.13.1* enumerates a few important QoS models.

As an example, we introduce the model from ITU-T Rec. I. 350. The essence of this model isthe definition of the 3 x 3 matrix approach for QoS and network

origin	model	characteristics
ITU-T	Rec.E.800, Rec.I.350, Rec.X.641, Rec.X.140, Rec.G.109	terms and definitions related to QoS/NP, QoS/NP in digital networks, Information Technology-QoS Framework QoS parameters for public data networks speech transmission quality
ETSI	ETR 003, ETR138, EG 201 769-1, Tiphon, ETS 300 416, TBR 21, TBR 38	Precursor of Rec.I.350 QoS indicators for ONP of voice telephony and ISDN, QoS parameters, definitions and measures, telecommunications and internet protocol
ISO/IEC	12326, 15802-3	corresponds to Rec.X.641
EURES-COM	P616, P806	quality classes, QoS/NP in multiprovider enviroment
IETF	RFC1633, RFC2475	QoS in intserv, QoS in diffserv

Table 1.13.1

performance, according to the *Table 1.13.2.*

The matrix defines nine parameters. These are the generic parameters and in addition primary parameters (to be determined on the basis of direct observations of events at service access points) depends on the service.

As an other example we show the structure of that part of TIPHON project, which deals with the quality of service, because this structure reflects the essence of this model:

- general considerations of QoS,
- determination of quality classes,
- control of quality of service,
- methods of QoS measurements,
- QoS tests,
- planning guide.

quality criterion communications function	speed	accuracy	Dependability
access			
user information transfer			
disengagament			

Table 1.13.2

1.13.4. Quality of service indicators [1.13.3]

According to the different QoS definitions, QoS is a non-quantitative term, but there are a lot of factors, characterizing QoS. The name of them is quality of service indicators or measures or determinant of QoS.They are measurable in objective way, with tools or, in subjective way, by customer opinion surveys. So, there are two groups of QoS indicators, objective and subjective ones. The objective measurements are performed generally by the service provider itself, hence these measures are named also as internal indicators. These are provider-oriented. On the other hand, the target of the subjective measurements is the customer. The subjective measurements consider the service from outside, so that the subjective indicators are external measures and customer-oriented.

The ETNO, on the base of 98/10/EC Directive, proposes the selection of QoS indicators as follows: QoS parameters should be easily understood by the public, and be useful and important to them,

- all parameters are applicable at the network termination point. Where measurements are possible, they should be made on the customer's premises, using in-service lines. To be as realistic as possible, real traffic rather than test calls should be used as a basis of the measurements, wherever possible,
- parameters should be capable of verification by independent organizations.This verification might be made by direct measurements or by audit of operator's measurements,
- the accuracy of QoS values should be set to a level consistent with measurement methods being as possible with costs as low as possible,
- the parameters are designed for both statistical and individual application. The statistical values should be derived by the application of a simple statistical function to the individual values. The statistical function should be specified in the standard. The standard should also contain guidelines on how statistically significant samples should be selected.

The QoS measures make a connection between the quality requirements and activites.

In the case of telecommunication services, ETNO lists the following telecommunications activities: provision, service support, repair, communication, billing and complain handling. The quality criteria of ETNO are: speed, accuracy, availability, reliability (these are objective quality criteria), security, simplicity and satisfaction. The last three ones are subjective indicators. At other actors, exist other determinants.

The abovementioned EU Directive proposes for the fixed telephone service the following set of quality indicators:

- supply time for initial connection,
- fault rate per access line,
- unsuccessful call ratio*,
- call set up time*,
- response times for operator services,
- response times for directory enquiry services,
- proportion of coin and card operated public pay-telephones in working order,
- bill correctness complaints.

QoS parameter definitions and measurement methods have to be taken into consideration on the base of ETSI EG 201 769-1 (2000-04).

The indicators, marked with star, are not obligatory. The above listed measures are in close connection with concession indicators, valid in Hungary (*Table 1.13.3)*. This table contains the target values valid in year 2001 and 2002, too. It is also worth to mention the comparable performance indicator system and databank of the QuEST Forum (see later), which contains practical QoS data of different telecommunications companies in statistically processed form. This system and its data are available for members of the Forum and enable them to have benchmarking for quality improvement purposes.

Comparable performance indicators: In the practice of UK service providers the comparable performance indicators are well known. The service providers have agreed in the frame of the Industry Forum, that they voluntarily offer some achieved QoS values at UK Regulator's (OFTEL) disposal for publishing. The aim of this process is to make known the actual variety of service quality to the customers in order that they should decide, which service provider is the best choice for them . This fact shows, that the QoS indicators have a certain role of consumer protection. From the other side, a comparable performance indicator system gives also a possibility for the service providers to compare the QoS of different services and carry out benchmarking with the aim to improve their own services. The name of elements in such indicator systems is CPI (comparable performance indicator).

It should be remarked that there exist further categories for the QoS indicators. For example:

- primary (measurable direct) parameters and
- secondary (derived from primary one) parameters.

  There exist generic parameters, too (see 1.13.3).

quality of service indicator	2001 [%]	2002 [%]
percentage of unsuccessful local test calls, initiated during the peak hours	1,40	1,40
percentage of unsuccessful national long-distance calls, initiated during the peak hours	2,80	2,70
percentage of unsuccessful international test calls, starting during the peak hours	2,80	2,70
percentage of test calls, directed to the operator service, answered within a specified time	97,00 T<20s	97,00 T<20s
percentage of test calls, without dial tone received	0,80 T>3s	0,80 T>3s
serveability of public payphones	92,00	92,00
monthly number of faulty telephon stations	0,02 nr/st/ month	0,02
annual average time of failure of subscriber stations	3,10 hr/st/year	3,10
ratio of repaired faults within 24 hours, related to the reported faults	0,90	0,90
ratio of billing complaints	0,016 nr/st/year	0,016

Table 1.13.3

telecommunications activities	objective QoS indicator	subjective QoS indicator
service provision	supply time for initial network connection, percentage of orders, completed on or before the date confirmed	⇐satisfaction with this parameter
restoration/repair	fault-clearing of reported faults on or before the date confirmed	⇐satisfaction with this parameter
service reliability	number of reported faults, number of customer touched by cut connection	⇐satisfaction with service reliability
billing	number of billing complaints received per 1000 units	⇐satisfaction with this parameter
complaint handling	% complaints, resolved within 20 working days	⇐satisfaction with this parameter

Table 1.13.4

1.13.5. Measuring of QoS [1.13.4]

Condition of service quality improvement is the measureability.The *Figure. 1.13.1* shows the principle of quality control, based on the duality of the quality measures (objectives and subjectives). Natural expectation is, that if a service is good, then the objective and subjective measurements both give good results . But when the service provided for customers is bad, then both the objective and subjective measurements are finished with faulty result. However there are significant

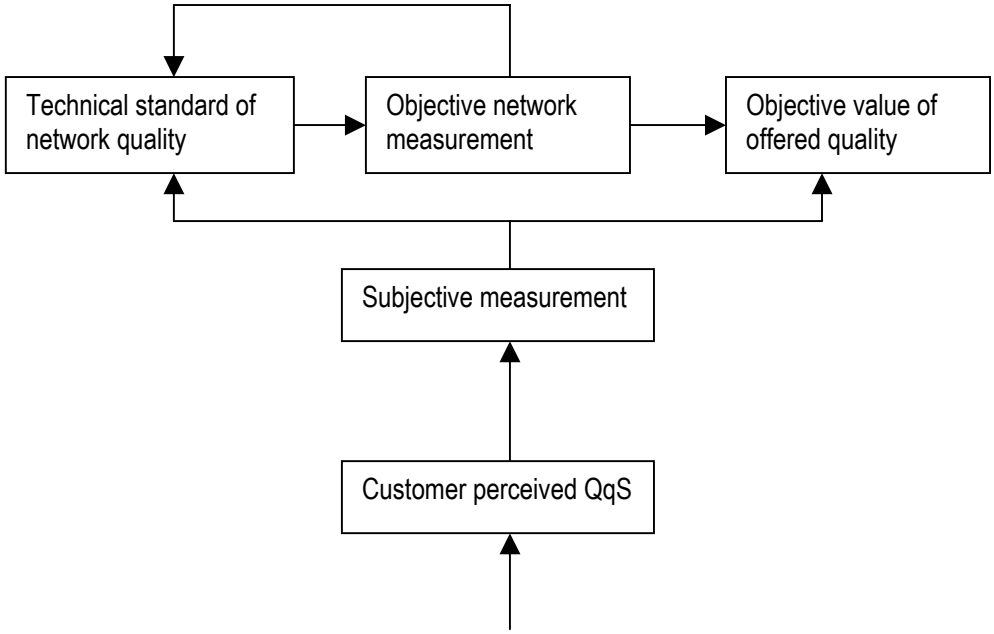


Figure 1.13.1

difference between the two type of measurements. If we want to make use of the advantages of both, the best is to combine the two methods, according to the *Figure. 1.13.2.*

On one hand, the QoS is measurable from the provider's point of view. In this case measurements are directed to determine network performance parameters. On the other hand, there are QoS measurements from the customer's aspect, where the end to end service quality measurements between customers are targeted. A lot of automatic measuring systems were developed for measuring network performances and QoS indicators.

Such systems consist of

- data collectors, connected to network elements,
- hardware and software controlling the data collection and
- central unit, processing the data.

1.13.6. Economical aspects of quality of service [1.13.5],[1.13.6],[1.13.7]

- Relationship of quality and cost:

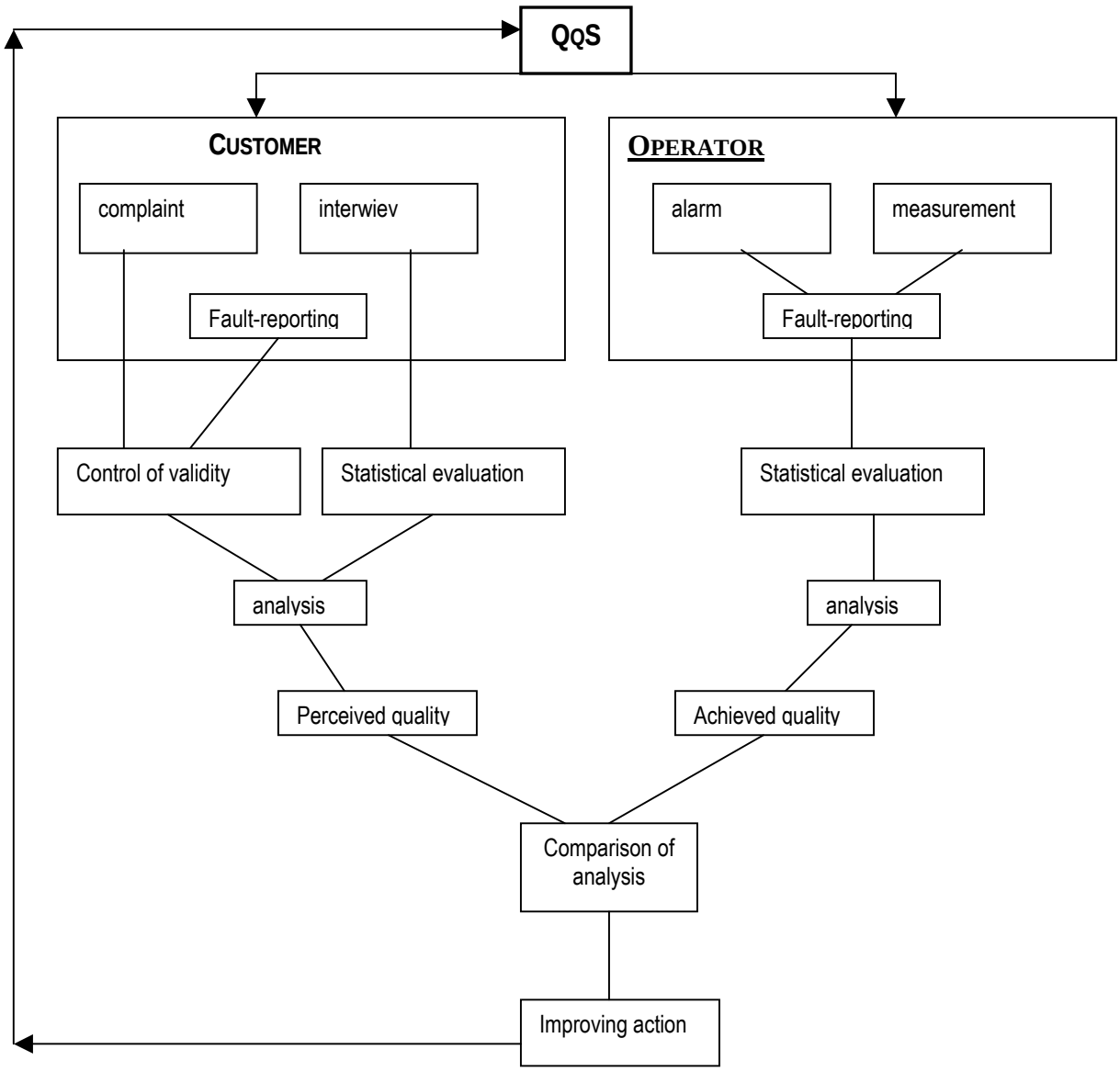


Figure 1.13.2

When a customer chooses a service provider, has requirements concerning the quality of service. The customer with carrier selection chooses a service provider, simultaneously taking into consideration his/her own experience from the past, the writing information, existing from the service, reputation and image of the service provider in the market and at last but not least the price of the service, too. The experiences and studies show, that there are close connection between the quality and price. It means, that the customer has a value-based selection, where the value is characterized by both quality and price. The costumer has to decide, how much can he pay for the good quality. In the traditional quality theory a rising curve gives the relationship between quality and cost. In the new approach, the cost

function includes the QoS, with introduction of a so called Q factor, as a variable. In the practice the Q factor means objectives (target values) of QoS indicators. Such a method has already been applied in the USA.

- Cost of quality (good and bad quality):

The business excellence of a company or an enterprise depends on some enablers, among them on the quality costs. The model of EFQM (European Foundation of Quality Mangement) reflects well the connection with quality costs.

The quality costs are divided into two groups: the costs of conformity and non-conformity. The sum of these cost-factors in a specific process represents the process-cost. The two cost factors move in opposite direction in the function of the quality as a variable. At a certain quality the process cost has its minimum value and this point represents the quality optimum. It is possible to interpret this partitioning on whatever level of any process or organization.

In the classic quality theory a curve with an absolute minimum, like a parabola describes the quality cost fuction. According to the newest researches, the optimum quality curve is a descending curve, like a hyperbola.

Another approach of the quality costs (BS 6143 model) distinguishes the following cost components: prevention cost, appraisal cost and (internal and external) failure costs. According to the experiences the ratios of them are approximately: 5:30:65.

In a project, within the frame of EU, a method for decreasing quality cost was elaborated. This method is applicable in practice.

- acceptability in the economic life, in the market:

This concept or measure is relevant on the market place and describes the attitude of users to technology and new technical application. Acceptability is in close connection with the quality of service. According to the definition it is the ratio of the number of people actually using the service and of the total number of people in the target group. Acceptability has

- service-specific,
- customer-specific,
- enterprise-specific and
- enviroment specific determinants.

The enterprise-specific determinants are independent of service. For example, such determinants are the organization itself, qualification of staff, intensity of competition, etc. Enviroment specific determinants are:

compatibility of service, accelerated cycle of innovation, infrastructure, advertisement, etc. Customer-specific determinants are: subjective advantages, motivation, expectations,, acceptability of risk prediction to technology, etc. At the end, service-specific determinants are: dimensioning, applicability, funtionality, charging, standardization of implementation and quality of service.

1.13.7. Regulation of quality of service [1.13.5], [1.13.8]

The key questions related to the QoS regulation are the following:

- character of the regulatory enviroment (monopoly or competitive enviroment),
- determination of QoS indicator set, published by service providers,
- audit of published data,
- policy applied to the service provider, as he/she does or does not perform the obligations (bonus or penalty),
- in case of new services, timeschedule of introducing QoS parameters,
- the method and steps of maintaining quality, by the service provider.
- In the era of monopoly, first of all in the 80-es there were two basic type of QoS regulation:
- *regulator-oriented type*, with characteristics:

  - service providers and/or regulators published the measured values of QoS indicators, in forms of reports or accounts,
  - regulators set target values for those indicators,

  in case the objectives are not performed, the service provider pays penalty,

- user-oriented regulation, with characteristics:

  - the service provider publishes achived values of QoS indicators,
  - the service provider offers determined values of QoS indicators, in contract,
  - in case the objectives are not performed, customer gets a compensation.

In the competitive environment it is necessary to change the regulation of QoS. A possible approach inside of the EU:

- the basis should be legal material of EU (for example 98/10/EC Directive) and ETSI standards, concerning the QoS indicators,
- another solution to be taken into consideration is a reference model, according to which

  - stakeholders of regulation are customers, service providers and regulators.

  Relationship of them is characterized by the

  - consumer loop (between customer and regulator),
  - market loop (between customer and service provider) and
  - regulatory loop (between service provider and regulator) and
  - key questions from the regulatory point of view: information, commitments, recourses.

  Table 1.13/5 shows a possible solution.

QoS regulation	information	commitments	recourses
market loop	proactive info about the performances, info about the faulty operation	individual commitments set by contracts	refund/compen-sation in case of QoS beyond targets, resolution of dispute
consumer loop	publication of information by NRA, when appropriate	-	resolution of dispute
regulatory loop	achiving of performance measurements # may be requested, # is obligatory	collective targets formulation of orders, monitoring weak parts	should be efficient tools in the hands of regulators against service providers (sanctions,too)

Table 1.13.5

1.13.8. International organizations [1.13.9]

A lot of international and european standardization organizations (ISO, IEC, ITU, ETSI, etc.), research centres (Eurescom, etc.) and organizations of stakeholders (manufacturers, service providers, customers) in telecommunication (IETF, INTUG, ETNO, QuEST Forum, etc.) deal with the general and special questions of quality of services. We introduce these organizations shortly, focussing to targets and activities in connection with the quality of services.

ISO (International Standards Organisation) was founded by 25 national standardization organizations, in year 1947. The aim of ISO was to publish international standards for helping the worldwide change of products and services. ISO does its job in close cooperation with IEC (International Electrotechnical Comission) and with ITU-T(International Telecommunications Union-Telecommunication Standardisation Sector). ISO issues some standards with same text, but with own numbering system. In the field of quality management the EN ISO 9000:2000 series is the most important.

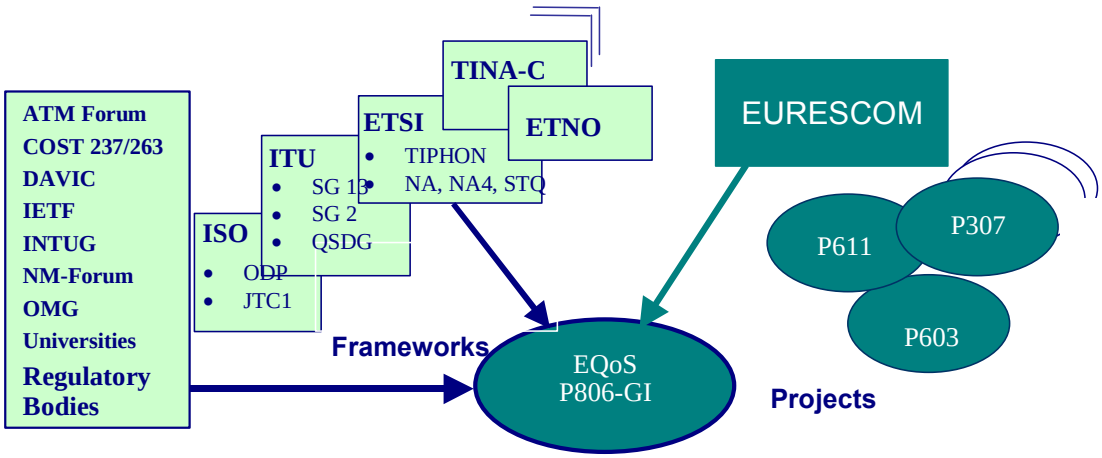


Figure 1.13.3

IEC TC 56 (Dependability) Technical Commitee deals with reliability questions of electrotechnical and electronic products and services. TC 56 issues standards, applicable well in the field of telecommunication. First of all, these standards deal with classification of failures, statistical data-processing and the organisation of maintenance. These items give a good help to the judgement of QoS.

ITU-T gathers the goverments, national telecommunication

administrations and recognized operating agencies, manufacturers, service providers, scientifical and industrial organizations and groups of consumer protection. The ITU-T does its job in study groups and working parties in those. Nowadays activities of five study groups are connected to quality of service. Altogether in the frame of 13 questions (Q5/2, Q1/7, Q2/7, Q8/7, Q7/12, Q9/12, Q13/12, Q15/12, Q6/13, Q7/13, Q8/13, Q9/13) are studied the quality of service. Results are issued in Recommendations. Recommendations dealt with QoS belong to E, G, I, M, Q, P, X and Y series. In addition the QoS is the main field of QSDG's interest. QSDG was created, inside SG2, in 1982. Its main aim is to improve the quality of international telecommunication services. Because of this, QSDG concentrates hundreds of experts of more than 100 countries and they exchange their practical and theoretical resultsregularly. QSDG deals with the all aspects of QoS.

The competence area of *ETSI* (European Telecommunications Standardization Institute) is Europe, but acceptability of their standards expands outside of Europe. Field of activities is standardization of telecommunication, including radiocommunication, broadcasting and information technology. ETSI was formed in the year 1988, creates technical standards and approves them. One part of its activity is performed in projects. Technical Commitee TC SPAN (Services and Protocol for Advanced Networks) which is successor to Network Aspects (NA) commitee, and the TC STQ (Speech Processing Transmission and Quality aspects) committee, as well as the TIPHONi project deal with qualiy of service.

Eurescom comprises researcher and developing activities of 20 european network operators in the frame of teamwork. Since 1991 has finished some usefull projects.

IETF (Internet Engineering Task Force) is an international community of engineers, operators, researchers, etc. IETF does the activities in connection with Internet, in working groups. Different working groups were created according to different topics. QoS is one of the different topics.

INTUG (International Telecommunications User's Group) acts in behalf the telecommunications users, among others in the field of quality of services, too.

ETNO (European Telecommunications Network Operators) is the forum of european public networks operators. It was created in consequence of the separation of operation and regulation. It has (among the others) a QoS Working Group, which has established a useful quality of service model, with the ETNO matrix representation. ETNO is important also from the following viewpoint: ETNO was the first from the european telecommunication organizations which initiated the unification of subjective QoS indicators.

In 1997, the leading telecommunications companies of the world, have founded a new organization, *QuEST Forum* (Quality Exellence for Suppliers of Telecommunication). The aim was

- to determine the special telecommunications requirements and measurement methods of qualification according to ISO 9000, summarized in a handbook TL 9000 (TL =Telecommunication Leadership),
- to train auditors for assessing and certifying the quality mangement systems of telecommunications manufacturers, operators and service providers on the base of TL9000 standard,
- to qualify and empower some registry organizations to accreditate certifying bodies.

TL 9000 is a multilevel system. The basis is the ISO 9000:2000. The general telecommunications requirements are built on the ISO 9000:2000. The following level are the requirements of hardware, software and services. And on the top there are the measures and measurement methods, needed for the control of quality requirements. TL 9000 Handbook consists of two volume. The first contains the quality requirements and the second the quality measures.

The QuEST Forum has more than 160 member enterprises in the world, in 2001. They send their QoS data regularly into the central databank of the Forum, where the data are statistically elaborated. Interpretation of these data is unified, so that they are comparable. Members of the Forum can use these data for both benchmarking and developing their own quality improvement programs.

1.13.9. Standards and projects, dealt with quality of service [1.13.10]

Standards:

- ISO standards:
  - MSZ EN ISO 9000: 2000 Quality management systems. Fundamentals and vocabulary
  - MSZ EN ISO 9001:2000 Quality management systems. Requirements.
  - MSZ EN ISO 9004:2000 Quality management systems. Guidelines for performance improvements
- ITU-T recommendations:
  - E series

checking the quality	(E.420-E.489)
QoS terms and definitions	(E:800-E.810)
QoS models	(E.810-E.845)
QoS objectives and planning	(E.845-E.899)

  - G series

general	(G.100-G.109)
telephone transmission quality	(G.110-G.119)
quality and availability target	(G.820-G.830)

  - I and P series

performance objectives	(I.350-I.359)
general network requirements	(I.370-I.399)
telephone transmission quality	(P.10-P.80)
quality assessment	(P.80-P.800)

audiovisual quality (P.900-…)

- M and Q series

measurement technik (M.1016-M.1040)

telephone transmission quality (Suppl. to M.series)

planning objectives Q.543

- X and Y series

data communications

IP based services

- ETSI standards:
  - EG 202 086 Objectives and principles for "traditional quality" telephone service
  - EG 201 377-1 Speech quality, Comparison performance methods,
  - EG 201 769, -769-1 QoS parameter definitions and measurement
  - EG 201 769-1 Parameters for voice telephony service

TR 101 329 TIPHON

TR 101 329-1 General aspects of QoS

TR 101 329-2 End to end QoS in TIPHON System

TR 101 329-5 QoS measurement methodologies

TR 101 329-6 Actual measurements of network and terminal characteristics and performance parameters in TIPHON networks and their influence on voice quality

TR 101 329-7 Guideline for planning

Projects:

- Eurescom projects
  - P 307 Reliability planning,
  - P 514 Network dependability
  - P 603 QoS measurement method selection
  - P 806 - GI A common framework for QoS/NP in a multiprovider environment
  - P 906 - GI Quality of service methodologies
  - P1008- PF Inter-operator interfaces for ensuring end to end QoS
  - P1003- PF IP QoS framework in case interconnection
- ITU-T projects
  - GII- N.8
  - IP-8
- ETSI projects
  - TIPHON,
  - …..

References:

[1.13.1] dr. A.V. Feigenbaum: Quality trends in the new Millennium. 44-th European Quality Congress, 2000.jun.

[1.13.2] Eurescom P906-GI project: QUASIMODO - Quality of service methodolgies 1999, www.eurescom.de

[1.13.3] dr. Buzás Ottó at all: Távközlési kultúra (Telecommunications culture)PressCon kiadó 2001

[1.13.4] dr. Veress Gábor at all : Minőségügyi jegyzet (Quality booklet) sorozat a Veszprémi Egyetem poszt-graduális oktatása számára, 1998.

[1.13.5] A.P. Oadan at all: Quality of service in telecommunications IEE Telecommunications series 39, 1997.

[1.13.6] BS 6143 Brithis Standard,

[1.13.7] Telecommunications Regulation Handbook World Bank, Washington 2000.

[1.13.8] Sagatel study for the EC. 2000

[1.13.9] www.iso.ch, www.itu.int, www.etsi.org, www.eurescom.de, www.ietf. org

[1.13.10] WWW.eurescom.de, www.etno.be

2. Transmission

In this Chapter, the scope of transmission is covered in two Sections. In Sec. 2.1, the realization methods of transmission paths are presented while in Sec.2.2, the transmission devices and methods of today's transmission technology are outlined.

In Section 2.1, the physical realizations of transmission methods for spanning distances are separately presented, including wireless and optical, further terrestrial and satellite microwave links. In these links, a variety of multiplexing equipment and modulation methods, together with bit error correction methods, are applied, all independent of the transmission path. The principles of these have already been presented in Secs. 1.3 and 1.5 so in the following, primarily technical aspects will be covered.

All these transmission techniques are included in a single Chapter, the two parts covered in Secs. 2.1 and 2.2 being strongly interrelated. They could have been treated in separate Chapters but editing requirements called for a single Chapter as the transmission equipment is heavily interrelated with the properties of the transmission path.

2.1. Transmission paths

In the first part of this Section, wireline transmission paths are presented. Sec. 2.1.1 covers the parameters characterizing transmission quality such as level, distortions, noises, stability, echo, bit error rate, jitter etc., further the effects of the modulation methods on the transmission quality. This is followed by surveying the structure and and main parameters of copper lines, coaxial cables and optical fibres in Secs. 2.1.2 to 2.1.4.

The second part of this Section is concerned with the radio channel which is the transmission medium of a wiresless system. Sec. 2.1.6 covers the characteristics of a radio channel (antennas, wave propagation, propagation models, time varying channel, fading effects, noises, disturbances, interference). This is followed by the describing of radio modulation systems in Sec. 2.1.7 (linear and nonlinear modulation methods, spread spectrum techniques, CDMA), covering further the properties of the

various modulation methods (bit error rate in channels without and with fading, synchronization requirements, sensitivity, nonlinear distortions). Finally, Sec.2.1.8 presents a special radio channel type, the wideband long distance microwave link (2 to 18 GHz), further the wideband satellite channel.

László Pap dr., Editor of the Chapter

2.1.1. Wireline transmission paths – from open wire lines to optical fibres

Sándor Gránásy, author

György Takács PhD, reviewer

Historically, the first telegraph and telephone transmission was realized by the open wire line, built along the path from elements delivered to the site such as columns, supports, isolators and cupper bronze wire. The limited number of lines constructed from these elements, further their sensitivity to external electrical and mechanical effects, necessitated the application of *symmmetrical cables* which have been manufactured in factories using special machinery. These cables have been covered by a metallic sheet, containing hundreds of conductor pairs or quads, and buried under the pavement or the border of the road if any. Subsequently, the laying of the cable and the connection of the cable sections, further the balancing to reduce crosstalk have been carried out in the field. Later, *coaxial cables* no longer required the supression of crosstalk. Finally, the number of repeaters have been reduced by the extremely low attenuation constant of *optical fibres* which offered total immunity to high voltages and crosstalk disturbances. Optical fibres still required the interconnection of manufactured sections but subject to terrain conditions, the number of interconnections can be significantly reduced. Summarizing, the problems arising in the field have essentially be solved within the factory. Naturally, the evolution previously outlined have made possible by the progress in other branches of wireline transmission system such as the appearance of amplifiers, the frequency division or time division multiplex systems, and finally the fibre components.

The efficient operation of any communication network requires from the outset the knowledge of requirements in the time and space domain. What kind of distribution network is required? In addition to the technological elements required for the realization, the probability of the number of the necessary intervention into the network, either for enlargement or for fault clearing, should be estimated, together with the methods to be applied for fault localization.

2.1.2. Principles of wireline connections

Sándor Gránásy, author

György Takács PhD, reviewer

There are two basic types of metallic conductors applied in wireline communication, the symmetrical and the asymmetrical type. Symmetry is meant with respect to ground. The former type includes open wire lines and symmetrical cables while the latter includes coaxial cables. In the former type, the degree of symmetry should be as high as possible and should be maintained. Coaxial pairs are in principle asymmetrical.

a) Transmission properties of symmetrical lines

In their application pairs are characterized by the image impedance $Z = [(R + j\omega L)/ (G + j\omega C)]^{1/2}$ and by the image (propagation) constant $\gamma = [(R + j\omega L).(G + j\omega C)]^{1/2}$,where R, L, G and C are the loop resistance [ohm/km], inductance [mH/km], conductance [1/ohm.km] and the effectiv capacitance [nF/km] per unit length, while ω is the angular frequency [rad/sec]. The calculation of frequency characteristics of these primary parameters is complicated and hardly useful. Their qualitative behaviour can easily be estimated: the capacitance is practically frequency independent, and the conductance is significant in the case of paper-air dielectric. For a polietilen air dielectric substance, capacitance and loop resistance increase slowly as a function of frequency. Inductance is constant, and following a slow decreasing interval, it is again constant as a function of frequency.

Practically, the measurement of secondary parameters such as the image impedance and image propagation constant is of importance. Four parameters can

be freely chosen: material and diameter of the conductor, material of a insulation and the effective capacity. The effective capacity can be influenced by the insulation material and construction (cordell, foam etc.), but choosing this later one must keep in mind to realize a mechanically stable structure. It thus follows that the parameters of symmetrical cables are chosen empirically.

In the audio frequency range, the above expressions can be simplified as follows:

$Z = [(R/j\omega C]^{1/2}$ and $\omega = [j\omega CR]^{1/2}$

The expression of image impedance and image propagation constant has identical real and imaginary parts. The real part of the image transmission has primary importance, being $\gamma = [\omega CR/2]^{1/2}$

The attenuation can be decreased witlh loading it means application of *concentrated inductivity coils..* In this case, the inductivity in the previous formulae will be significant, resulting in $Z = [(L/C]^{1/2}$ and $\gamma = [R/2Z]$. The line thus produced is acting as a lowpass filter, and has a high attenuation (more than 15-25 dB) the transmission above the audio frequency range.

In the carrier frequency range, the image (characteristic) impedance varies according to the (-1/2) power of frequency. This change is not significant above 12 kHz, and tends to a constant value between 150 and 170 ohm, dependent on the dimensions and the insulation material. The group delay (phase constant) decreases continuously so the characteristic impedance can practically be regarded as ohmic. The frequency response of the propagation constant changes according to the ½ power. In this frequency range, the insulation material has a significant effect. This can be demonstrated by comparison of different plastics. .Both with different conductance and dielectric constant have a significant role: approximately identical attenuations have been measured by following two lines: copper wire of 1.2 mm diameter, with paper cordel and ribbon insulation at 252 kHz on one hand, and copper wire of 1.3 mm diameter, with stiroflex cordel at 552 kHz, on the other hand. The former was manufactured with an effective capacity of 26 nF/km, the latter with 23 nF/km, which seems to be an insignificant difference, however, the difference between high frequency losses is significant. Subsequently, the polyetilene, having

favourable electrical properties, has been extensively applied as the basic insulator in the cable industry.

b) Crosstalk properties of symmetrical cables

Within a symmetrical cable, the pairs of any quad have an influence on the pairs of each other pairs of the cable quads. The number of possible relations is quite high, even if only adjacent pairs are taken into account. To provide crosstalk equalization, coupling has to be determined, but for other purposes, it is sufficient to measure the crosstalk attenuation. Crosstalk attenuation is the logarithm of the ratio of undesired and useful signals. The numerator of this ratio is the power in the interfering circuit, $P_1 = U^2/z_1$ while the denominator is the power measured in the interfered circuit, $P_2 = V^2/z_2$. (Generally, the two impedances may be different but in most cases, they are identical.) The crosstalk at the feeding end of the circuit is called near-end crosstalk while at the receiving side it is called far-end crosstalk. Crosstalk is introduced by the coupling impedance or admittance between the interfering and interfered circuit, the magnitude and the phase of this coupling is changing randomly. The reasultant of these elementary couplings can be defected at both ends of the disturbed ciruit. Any elementary coupling is introduced by a voltage source having a longitudinal impedance with zero internal impedance, terminated in the input and output direction by the image impedance of the circuit. This means that half of the power is propagated towards the input, and half of it toward the output. The interfering signal detected at the near end is due to following effects: it is attenuated according to the propagation value of the interfering circuit up to the point of the elementary coupling. Finally, its value is decreased, from the coupling point to the input of the circuit, according to the interfering circuit image attenuation. The crosstalk attenuation at the input, created by an elementary coupling at a distance x from the input, is increasing by an attenuation of $2\alpha x$ where α is the real part of the propagation constant $\gamma = \alpha + j\beta$ and $\beta$is the phase constant. From the foregoings, two conclusions may be drawn.

1) The near-end crosstalk can be compensated *only at the point of origin*, by inserting an element of equal mangitude into the opposite relation, 2) at the two ends of the repeater section, it is feasible to apply cable sections having the best near-end crosstalk.

The above described balancing method should be applied also for the far-end crosstalk: the balancing capacities and resistance will be also inserted at the coupling points. However, the interfering signal due to far-end crosstal decreases with the attenuation between the coupling point and the terminal point of the interfered circuit. The signals due to elementary couplings along the line give the resultant value by vectorial addition.

Near-end and far-end crosstalk appear simultaneously, and the type of application will govern their relative importance and the degree of interference which can be accepted. Quality is determined by the *signal-to-noise ratio [dB],* the ratio of useful to disturbing signal powers, which should exceed a threshold set by the modulation system. This condition is sometimes characterized by the *crosstalk protection.* The useful signal arrives (over the interfered circuit) at a level decreased by the link attenuation *a*: (P-*a*) db. The level of the disturbing ciricuit is also P. The propagation directions of the disturbing and disturb signals are opposite. In order to obtain the required signal-to-noise ratio, the crosstalk attenuation should exceed the given (recommended) minimum level of signal-to-noise ratio increased by the link attenuation. This is the *near-end crosstalk requirement.*

The signal-to-noise ratio should be investigated at the far end if the propagation directions of the signal of both (disturbing and disturbed) are identical. By definition, the far-end crosstalk attenuation is the difference between the levels of the original (disturbing) signal at the input and the disturbed signal at the far end. From this, the signal-to-noise ratio or the far-end crosstalk protection is obtailned by decreasing the far end crosstalk attenuation by the link attenuation.

Based on the foregoings, it is feasible to define the crosstalk phenomenon in a more general form: high-level circuits may introduce disturbances into low-level circuits. The repeater station input terminals carry always low-level signals while the output terminals are at high level. It may happen that some wires in the cable are not interrupted for inserting amplification. In this case, the amplified signal may reach the non-amplified side by double coupling. This kind of crosstalk is known as *double near-end crosstalk or interaction crosstalk.* This problem is encountered in single cable (the two propagation direction of carrier systems are transmitted in the same cable) systems equalized for carrier frequeny transmission (Figure 2.1.2.1), further in

all two-cable systems comprising voice frequency quads in addition to carrier frequency quads, in at least one of the cables.

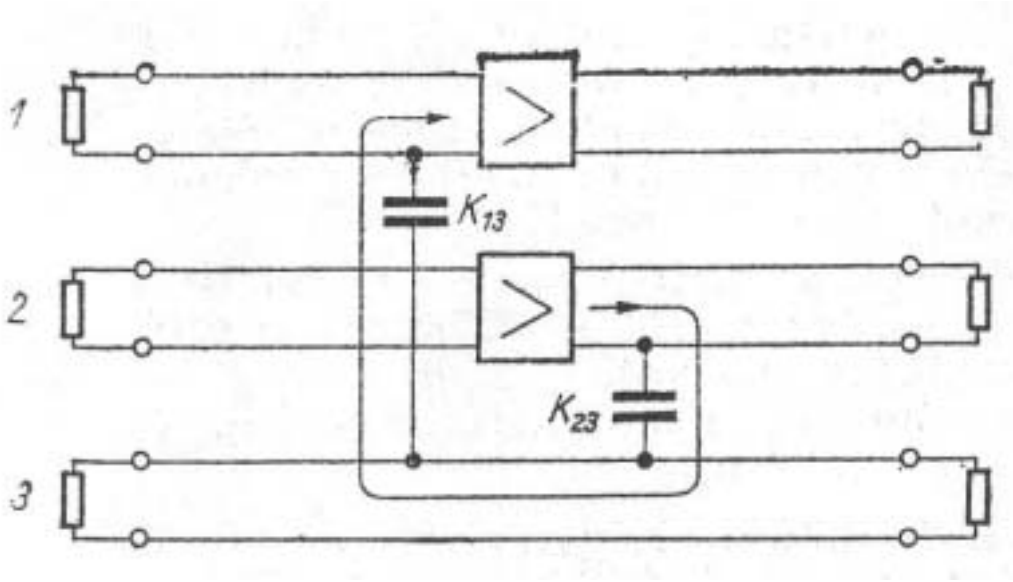


Figure 2.1.2.1. Illustration of double near-end crosstalk

This is a type of near-end crosstalk, so between any output and input terminal, a signal-to-noise ratio which is higher by the section attenuation should at least be obtained. In two-cable systems, it is sufficient of interchange the roles of cable *A* and cable *B* (Figure 2.1.2.2).

In this way, one of the cables will have the low-level signals in a given point while the other on the same cross-section will include the high level signals. In single cable systems, longitudinal bifilar wound coils should be inserted which will increase

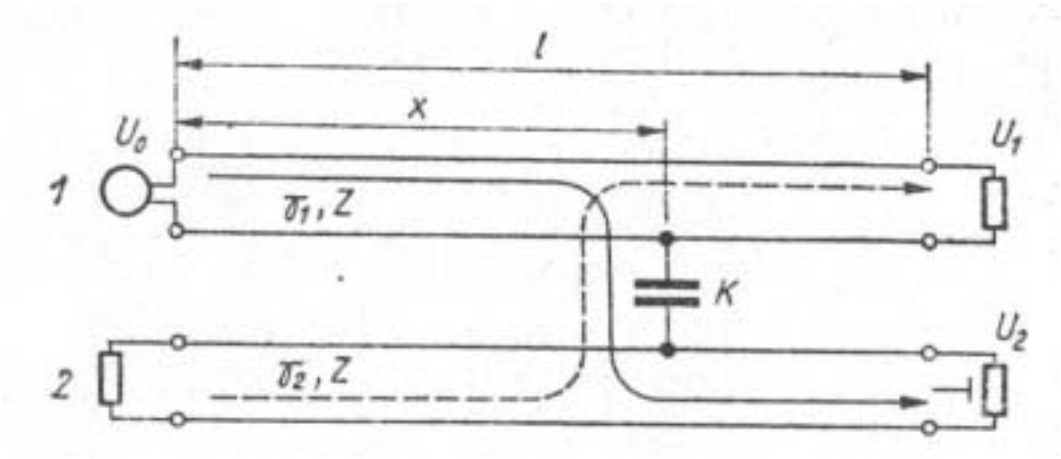


Figure 2.1.2.2. Illustration of double near-end crosstalk

the crosstalk attenuation without influencing the original circuit function (Figure 2.1.2.3).

Crosstalk attenuation appears as the complex vectorial resultant of the partial couplings. These couplings are of statistical nature due to manufacturing tolerances. *Crosstalk attenuation has a tendency of decreasing as a function of frequency.* Crosstalk attenuation is a cable parameter that can somewhat be increased by the insertion of balancing elements. However, both the cable type and the manufacturing and balancing technology have restrictions: paper-air insulation allows the transmission of 60 channels (252 kHz) over 7 quads in 14 phisical circuits while stiroflex or polyethylene-air insulation allows the transmission of 120 channels (552 kHz) over 7 quads in 14 phisical circuits. The requirement concerning near-end crosstalk between the two directions could be achieved only by two-cable systems.

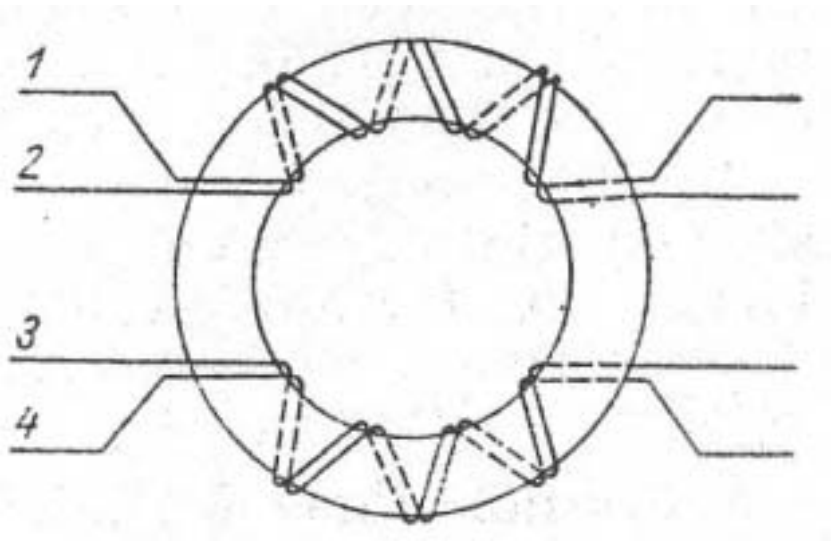


Figure 2.1.2.3. Wiring diagram of a longitudinal coil

c) Transmission parameters of coaxial pairs

In the practically applied frequency range, the main transmission characteristics of coaxial pairs are independent from external circumstances or influences. The frequiency range below 60 kHz is not applicable due to crosstalk. Primary parameters of coaxial pairs can be expressed by formulas comprising the simplified form of higher order Bessel functions, and their validity is proved by measurements on symmetrical cables. The formulas related to image impedance and image propagation are valid for coaxial cables too. The serial member is the sum of

the external and internal conductor serial impedance and the reactance of their inductive loop:

$$(R + j\omega L) = (1 + j).[1/2a + 1/2b].(\omega f/\omega\omega)^{1/2} + jf\omega.\ln(b/a)$$

where *a* is the radius of the inner conductor, *b* is the internal radius of the outer conductor (tube), $\omega = 1{,}26.10^{-6}$ [Vs/Am] is the permeability ($\omega_r = 1$), and ω[1/ohm.m] is the conductance of the outer and inner conductor.

G is negligible in the parallel member of the equivalent fourpole. The admittance is governed only be the cylindrical condenser formed by the two conductors. Its value is frequency independent:

$C = 2\omega\omega/\ln(b/a)$, where $\omega = \omega_0\ \omega_r$ and $\omega_0 = 10^{-12}$ [Vs/Am], and ω_r is the dielectric constant of the material between the two conductors. Its relative value is between 1 and 2, dependent of the ratio of polyetilene and air which are most frequently used. In view of the above relations, the wave impedance and the wave transmission measure can be determined:

$$Z_0 = (1/2\omega).(\omega/\omega)^{1/2} \ln(b/a) + (1 - j).(1/4\omega)[1/2a + 1/2b].1/(\omega f\omega\omega)^{1/2}$$

and

$$\omega = j\omega (\omega.\omega)^{1/2} + (1 + j) [(1/2a + 1/2b)/\ln(b/a)](\omega f\omega^{1/2}$$

It should be noted that the image impedance has two part. The first one is independent term Z_ω and the second a frequency dependent term. This latter is varying as the (-1/2) power of frequency. In coaxial pairs used in telecommunication with 2.6/9.5 and 1.2/4.4 mm nominal diametres, the real part of the image impedance at 2.5 and 1 MHz is 75 ohm. The image transmission is increasing with the (1/2) power of frequency. With outer and inner conductors of the same material, the attenuation has a minimum value at b/a=3.6. This is not a sharp minimum value which would be compulsory for the designer.

For operating coaxial pairs, the reflection has to be considered. The value of reflection is the ratio of the reflected and input signals, with or without the correction due to attenuation. Transmission performance is impaired by reflection, especially in case of picture transmission. Reflection is due to impedance irregularities along the length of the coaxial pair, caused by the manufacturing inhomogenity. Reflection can

be also the consequence of impedance difference at joining two factory length of cable.

A part of the signal is reflected to the input, and then is pursuing the original signal. This has no substantial effect because of the time distribution of the signal., if the propagation time is less than the recommended value. Transmission performance may severly be impaired because of following situation. At the joint, the signal is reflected towards the input at time instant t. Subsequently, it is again reflected at joint (n-1) at time instant (t+τ). However, it then follows the signal, and arrives at joint n at time instant (t+2τ), and arrives at joint (n+1) at time instant (t+3τ). It is reflected at joint (n+1) at time instant (t+ω), and arrives at joint n where it is again reflected, and arrives at time instant (t+3τ) at joint (n+1). It is thus seen that the signals originating from double reflection are added as complex vectors (Figure 2.1.2.4), producing echos, noise, distortion or, with picture transmission, a shifted disturbing second picture.

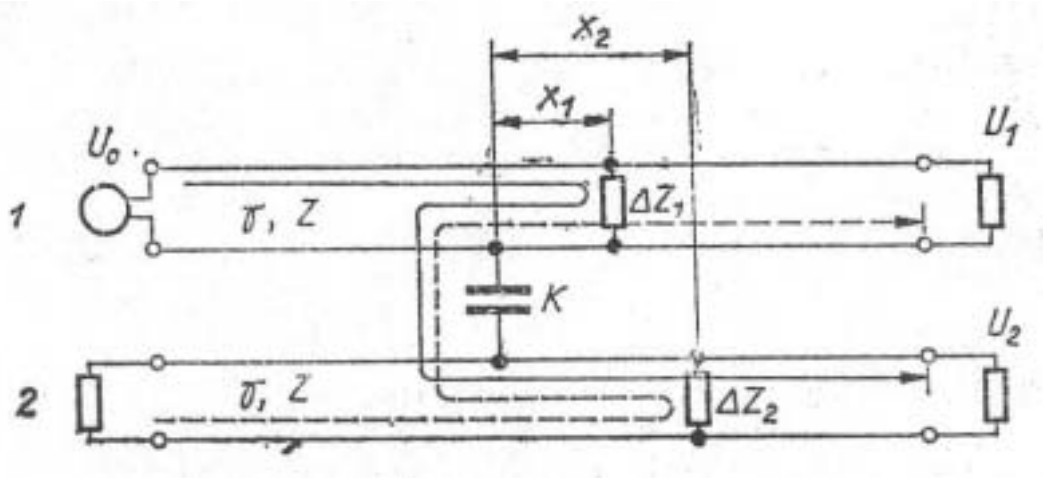


Figure 2.1.2.4. Illustration of reflected near-end crosstalk

To avoid this situation, the effective impedance differences between consecutive cable sections should be as low as possible. To achieve this, the following method proved to be the most effective. From the image impedances of the factory length, hystograms are prepared, and the area of the hystograms have to be divided into groups of quasi equal impadences, the number of groups being equal to the number of coaxial pairs in the cable. Into the individual cable sections, one coaxial pair from each group has to be installed. In this way, the possible impedance

differences are decreased by a fraction which is equal to the number of groups. Finally, in cascading the cable sections, the sections having nearly identical impedances should be joined.

d) Crosstalk between coaxial pairs

The outer conductor of coaxial cables is covered with a double-layer shield made of steel band. In addition to the shielding effect, this provides complete closure from end to end. The efficiency is further enhanced by copper coating of the steel band. This structure provides high enough crosstalk attenuation between two coaxial pairs so that in a signle cable system, the near-end crosstalk requirement between opposite directions will be met with safety.

For applications in metropolitan networks, mini-coaxial cables with 0.8/2.2 dimensions have been developed for links between exchanges, utilizing secondary or tertiary PCM transmission. PCM transmission requires less crosstalk attenuation, and this could be achieved with a shield having only a single steel layer.

Two coaxial pairs serve a multichannel multiplex system.

In the field of wideband communications with metallic conductors, coaxial cables are regarded as first class devices. However, they cannot be taken into account for new toll circuit, where the glass fiber is preferable.

2.1.3. Tools and technology applied in wireline transmission systems

Sándor Gránásy, author

György Takács PhD, reviewer

In *open wire cupper, bronze lines* of 2 to 5 mm diameter are applied. Bronze represents a compromise for meeting simultaneously the requirements of low resistance and greatest possible tensile strength. The bronze wire has been fastened to insulators mounted on steel supports, attached to wooden poles. The nominal distance between poles is 50 m. The wires have been fastened with hang-down provided for taking into account temperature changes. The different pole

arrangement can be seen on Figure 2.1.3.1. The different schemes are offering other coupling between the circuits, but the price of the less coupling is the worse usage of the pole.

The arising crosstalk problems have been solved by applying transposition. For minimizing the crosstalk between opposite directions, different frequency ranges have been alotted for the two directions, and so-called separate frequency n+n systems have been applied.

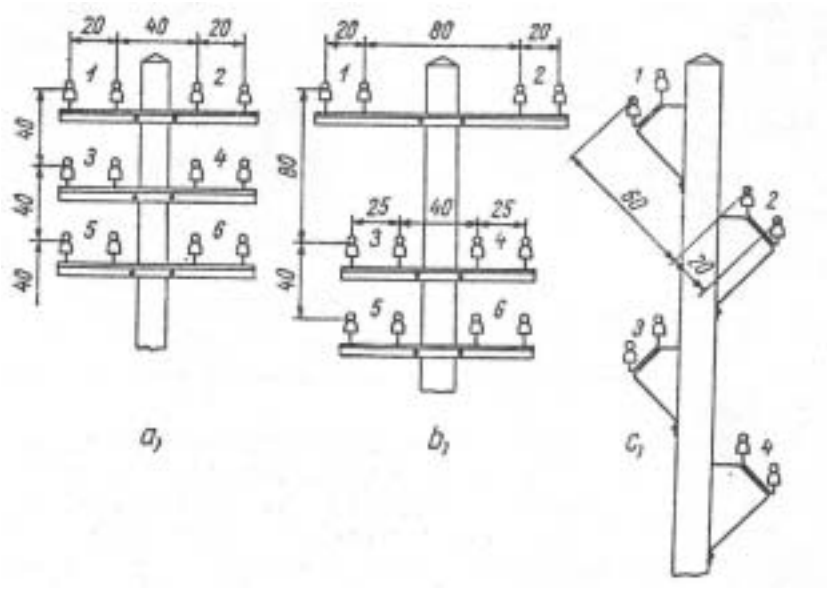


Figure 2.1.3.1. Poles applied in toll circuits for transmission of voice and carrier frequency signals

The pole-route holding open wire lines is extremely vulnerable, both mechanically and electrically. High coupling is introduced by the distance of more then 20 cm between wires and less 40 cm between the circuits, in the case of long haul circuits. In addition, the swiching and transmission equipment has also to be protected against over voltag and currents generated by high power long-distance transmission lines and atmospheric discharges. An extensive thunderstorm could have caused mass hystserics. Further, the number of circuits that could be installed on a series of poles holding open wire lines is limited, and the network itself is extremely sensitive fo external effects.

The first *cables* have been manufactured with *copper conductors* and lead sheath. The copper conductor has been isolated by a paper band, and the wires thus

obtained have been twisted to form pairs or quads. This can be regarded as the basic cable type which, in the course of time, has been followed by several variants. The use a differing pitch in adjacent pairs or quads may partly or completely solve the crosstalk problem. Further, the twisting will also provide suitable mechanical stability.

The conductor is made from copper or *aluminium*. Howeever, the latter can be applied only as a stopgap arrangement, in spite of its favourable electrical properties and price. In the following, only copper conductors will be considered. *The conductor diameter* is 0.4, 0.6 and 0.8 mm for local cables, 0.9 and 1.3 mm for long distance cables, and 1.2 and 1.3 mm for carrier frequency long distance cables. *Star quads* are applied for local and carrier frequency long-distance cables, and *DM quads* are applied for voice frequency long-distance cables (Figure 2.1.3.2),

In the star quad, wires are placed along the vertices of a fictive square, and circuits are made up from the diagonally placed wires. In the DM quad, two pairs with different pitch are twisted to form a quad. The pair as a circuit has a characteristic effective capacity, which has a nominal value of 38-56 nF/km for local cables, 38,5 nF/km for voice frequency long-distance cables, and 26,5 or 28 nF/km for carrier frequency toll cables. There are two methods of cable shaping: using either concentric layers of quads, or using braided groups of a great number of pairs (Figure 2.1.3.3)

As a first step, a cornice is formed by twisting one or more quads. This is the *core* of the cable, and this is enlarged, also by twisting, by one or more layers. The

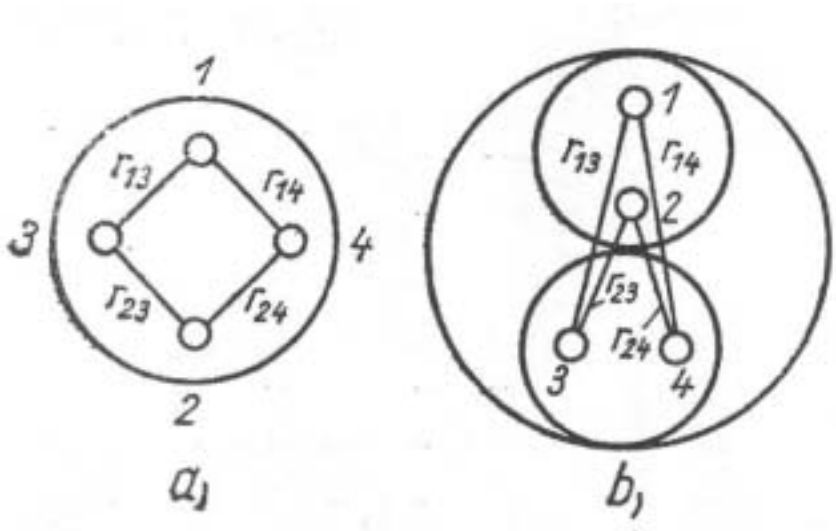


Figure 2.1.3.2. Cross-section of quads. a) star quad, b) DM quad

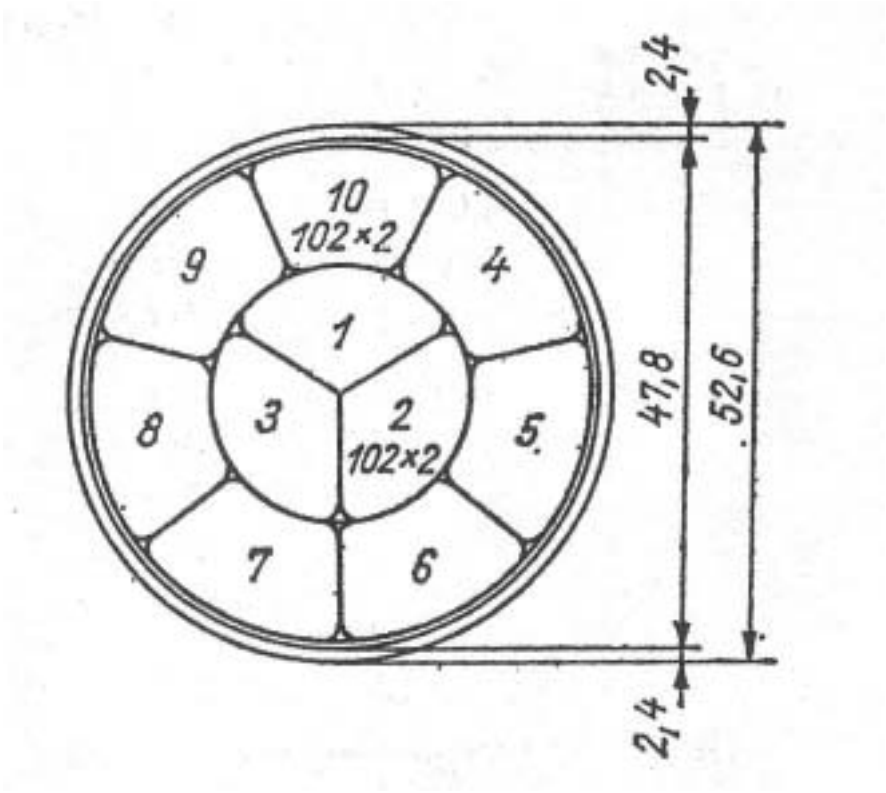


Figure 2.1.3.3. Cross section of a cable using braided groups

braided structure is justified by two reasons: either the capacity limitation of the twisting machine or the division of the quads into suitable units, so the braids can more easily be decomposed into branching cables. The twisted structure thus formed is covered by several layers of paper bands, called *belt insulation.* This is called cable core, and by pressing, it is covered by a lead sheath. The pull-in cable is thus produced. Excluding rare exceptions, the bare lead sheath can also be used as a pull-in cable. In the last 30 years plastic take over the role of paper forming a plastic-air insulation and also the lead is no more necessary by using different plastics sheats (more detail see on the next page).

In earlier applications, it was required to place the cable directly into the soil, i.e. into the working ditch. This required additional cable layers to avoid corrosion and to provide mechanical protection. These layers are the following: *cushion layer* made up of multi-layer paper band soaked in bitumen, further the *double steel band* and a jute layer soaked in tar, i.e. the *armor,* thus forming the armored cable. Later on, the steal band proved to be also a useful protection against electrical interference.

The requirement to utilize the space more economically and to provide efficient co-operation has been raised by the practice of laying cables along the same

path at different times, further the necessity of sharing the available pavement cross-section with the utilities. This led to the establishment of the telecommunication underground structure (ducts) and the coordination of the participants.

The first underground structure (ducts) utilized pre-fabricated monolithic concrete elements with several tube openings. These are at present substituted by plastic tubes piled up at the spot. Man-holes are used to facilitate the cable pull-in procedure and the interconnecion of the individual cable sections. Neither the concrete nor the plastic underground network can be made *watertight* by rational investment.

The underground structure represents a significant value, and due to the capacity of the installed cables, the requirements of several users can be met only with difficulties. This means that the available cross-section should be economically utilized. The underbuildings and the cables in it may be endangered by the *high power machinery* used for civil engineering. Only a suitable coordination can provide sufficient protection. This kind of complex task can be handled by a single corporation only so the operation and maintenance should be entrusted to a single company.

Neither the direct cable laying into the ground nor the pull-in procedure can provide an optimal solution for all situations. By necessity, the overhead cable has been developed, produced by connecting continuously or section-wise the pull-in cable and the steel supporting rope. In addition to these three solutions, the *cable duct* and the *utility tunnel* should be mentioned. The former provides *continuous acces*, e.g. in the manufacturing plant, and protection against mechanical effects. The latter is provided for several utilities.

A special cable class is formed by armored cables used for river crossings and for undersea applications, distinguised by the armor properties. As a result of growing pulling strength, the armor elements will be tightened more and more to each other, thus preventing the damaging of the cable by the penetration of sharp substances.

A new area in cable manufacturing has been marked by the utilization of plastics, primarily polyethylene. As a result of many years of research, the *skinned cable* with *vaseline filling* and *polyethylene foam insulation* has been developed. The foam has the effect of decreasing the dielectric constant and thus the cable

dimensions. The skinning process produces a smooth and closed surface, thus preventing the penetration of the vaseline into the polyethylene layer, and thereby the growth of the dielectric constant. The vaseline filling is intended to completely fill the residual space. The twisted wires are covered by several layers of polyethylene bands and finally by the aluminium shield. The polyethylene sheath is prepared by an extrudation process.

The above procedure results in a perfect cable which is significant in two respects.

a) A possible damage of the sheath, leading to a complete leakage with cables having paper insulation and lead covering, may remain unnoticed in the case of cables with vaseline filling.

b) The substitution of the lead covering with polyethylene covering has the effect of decreasing the cable weight, a useful feature during the transportation and the pull-in procedure.

Considering overhead cables, the figure-8 shaped self supporting cable is the most suitable. Some of these types have a cable core which has a somewhat wavy shape, resulting in a slight increase of the cable length. Cutting up the cable sheath between poles allows then to produce a subscriber branching. The main application field of this method is in the rural distribution or subscriber networks. This method is also applicable in rocky areas for establishing optical fibre toll circulits, provided the fibres will sufficiently tolerate the bending forces (Figure 2.1.3.4).

The *aluminium sheath* without a protective layer is unsuitable for pulling in because of corrosion. Further, the rigidity introduces bending difficulties. This

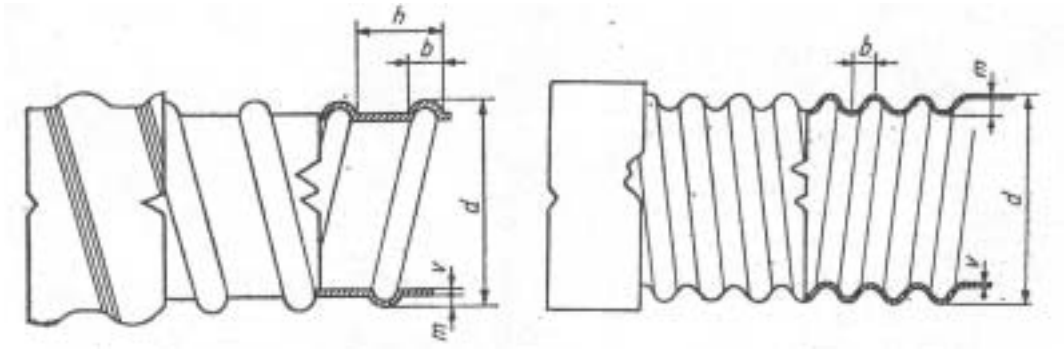


Figure 2.1.3.4. Wavy aluminium and steel sheath

problem could be mitigated by applying a wavy cable sheath but this would result in larger dimensions which were not acceptable. However, for meeting special requirements, cables with aluminium sheath could be used to fabricate armored cables. These are applied for excluding electrical interference by their high protecion capability.

Twisting metallic surfaces wilthout soldering results in a wire joint having a contact resistance with uncertain magnitude and stability. Therefore, this is not an acceptable solution, especially for data transmission. At new establishments, principally utilizing cables with vaseline filling, wire joints should be prepared with joining sleeve elements or joining moulds which have the equivalent contact resistance.

Various devices are used by manufacturers to provide symmetry between the outer and inner conductor of coaxial pairs: by equidistant discs, by moulds resembling to bamboo or sausage or double spiral. The individual solutions can electrically be exchanged. Two methods are used to interconnect coaxial pairs, depending whether special auxiliary devices are used or not. In the former case, the inner conductors are soldered by using a splitted sleeve, while the outer conductors are soldered with the aid of two half-sleeves. At the joint, the dielectric material has to be replaced. According to the original ratio, the diameter of the element connected to the outer conductor will be increased. The continuity of the steel bands is provided by a pair of steel sleeves. In the second case, only the silver band used for welding can be regarded as external material. A piece of silver band has to be inserted between the two inner conductors. Welding is followed by grinding off the excess material. The two tube-like outer conductors are first straightened, and then cut so as to contact each other. The contacted surfaces are then covered with a silver band of appr. 2 mm width which is folded back at the ends. The welding process is then carried out, and the insulation is replaced in original form. The outer conductor is then again formed by a mould to obtain a tube-like shape. Finally, the steel bands are replaced and fastened.

Chapters 2.1.1., 2.1.2., 2.1.3. translated by Tamás Sárkány dr.

2.1.4. Optical fibres

Gyula Veszely dr., author

György Lajtha dr., reviewer

In this paragraph the following notations will be used frequently:

The phase constant of a plane wave in vacuo: $k_0 = \omega\sqrt{\varepsilon_0\mu_0} = 2\pi / \lambda$.

The phase constant of the fibre in the axial direction: β.

The core index: n_m, the cladding index: n_h.

2.1.4.1. Types of optical fibres

The elementary lightguide or optical fibre is a dielectric cylinder surrounded by an annular dielectric cylinder of a little smaller refractive index. The inner dielectric is the *core* the outer one is the *cladding.* The step index (SI) fibre has a core with constant refractive index, the graded index (GI) fibre has a core with varying index.

In the case of SI fibres the Maxwell's equations can be solved analytically. When the mode is a *propagating* one, then its field as a function of the radial coordinate is oscillating in the core and exponential-decaying-like in the cladding. (As a consequence of this tail of the field, penetrating into the cladding, there is a crosstalk between fibres being too near each other). Decreasing the frequency the exponential-like function becomes more and more flat, while at a given frequency it becomes constant, i.e. the field in the cladding is the same at every radial coordinate. In such a case the fibre begins to radiate in the radial direction. The frequency at which this phenomenon occurs is called the *cut-off frequency* of the given mode. It is obvious that the fibre is unsuitable for transmission already in the vicinity of the cut-off frequency, because the field has to have infinite space to remain unperturbed. At the cut-off frequency the phase constant of the mode $\beta = k_0 n_h$ i.e. is the same as the phase constant of a plane wave propagating in the material of the cladding.

Increasing the frequency the slope of the exponential-like function becomes more and more steep, while at infinite frequency the field is totally withdrawn inside the core and will be zero in the cladding. In this case the phase constant of the mode

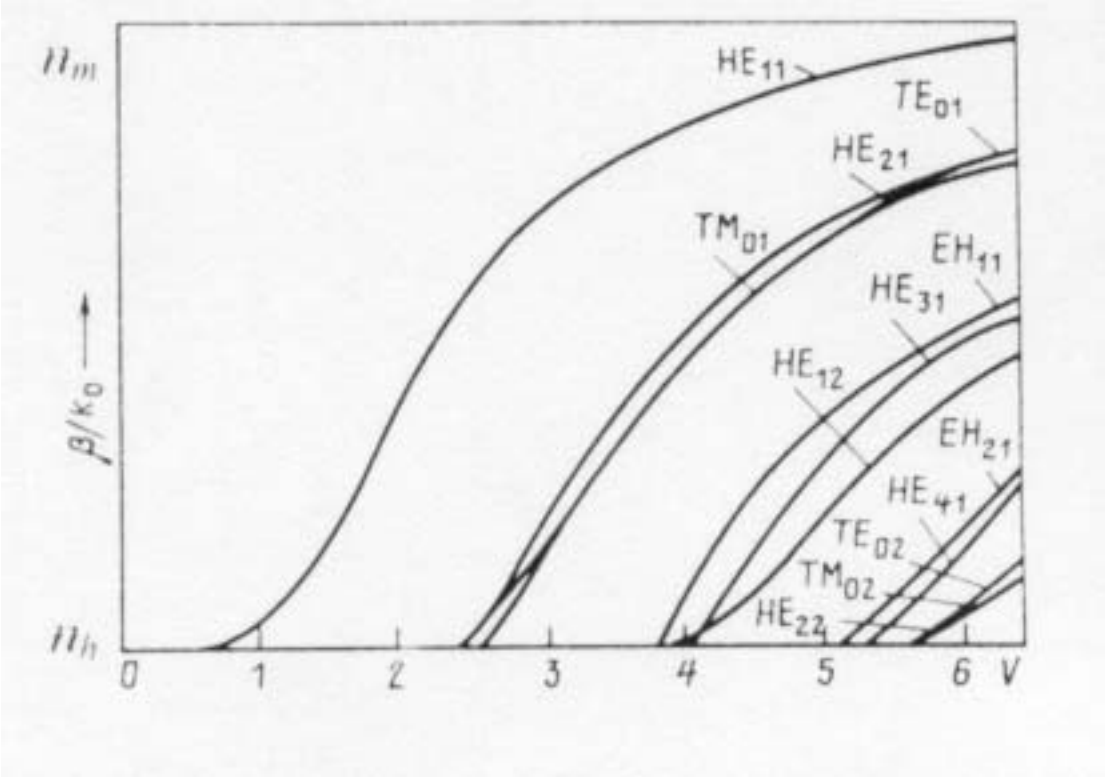


Figure 2.1.4.1. Dispersion curves of a SI fibre

$\beta = k_0 n_m$, which is obvious, because the field experiences only the index n_m. According to the aforesaid the dispersion curves can be seen in Figure 2.1.4.1.

The most important characteristic of the fibres is the V parameter (normalized frequency, normalized radius, mode-volume). The V parameter for SI fibre is

$$V = k_0 r_m \sqrt{n_m - n_h} \qquad (2.1.4.1)$$

where r_0 is the radius of the core. As can be seen from Figure 2.1.4.1. too the condition of the single-mode (SM) propagation

$$V<2,4 \qquad (2.1.4.2)$$

As the wavelength of the optical sources and the refractive indices are given upon other considerations (2.1.4.2) can be fulfilled choosing an appropriate small core radius. The typical core radius of the single mode fibres is 5-8 μm.

The typical core radius of the multi-mode (MM) fibres, which are less sensitive for the fabrication tolerances, is 50-100 μm. The MM fibres are fabricated with continuously varying core index. The reason is as follows.

Let us consider an impulse, AM modulating the light-carrier of frequency ω_v. The bandwidth of the impulse is small enough to approximate the dispersion curve $\beta(\omega)$ in the environment of ω_v by a straight line. In this case the modulating signal propagates with the group velocity

$$v_{cs} = \left.\frac{\partial\omega}{\partial\beta}\right|_{\omega_v} \qquad (2.1.4.3)$$

As can be seen from Figure 2.1.4.1. this group velocity is different for the different modes. That is, if we entrust the transport of the impulse to many modes, than the impulse carried by different modes will be arrived at the end of the fibre at different moments: the impulse will be widened. This phenomenon is called *intermodal* dispersion.

As can be seen fro Figure 2.1.4.1. the group velocity of the higher modes is smaller. At the same time the field of the higher modes are concentrated at the boundary of the core. The decreasing of the group velocity can be compensated by the lowering of the core index toward the boundary of the core. So the essential part of the field experiences a lower index, i.e. the group velocity increases. This the reason of making the MM fibres with index of refraction monotonically decreasing from the center of the core. In Figure 2.1.4.2. power-law index profiles can be seen.

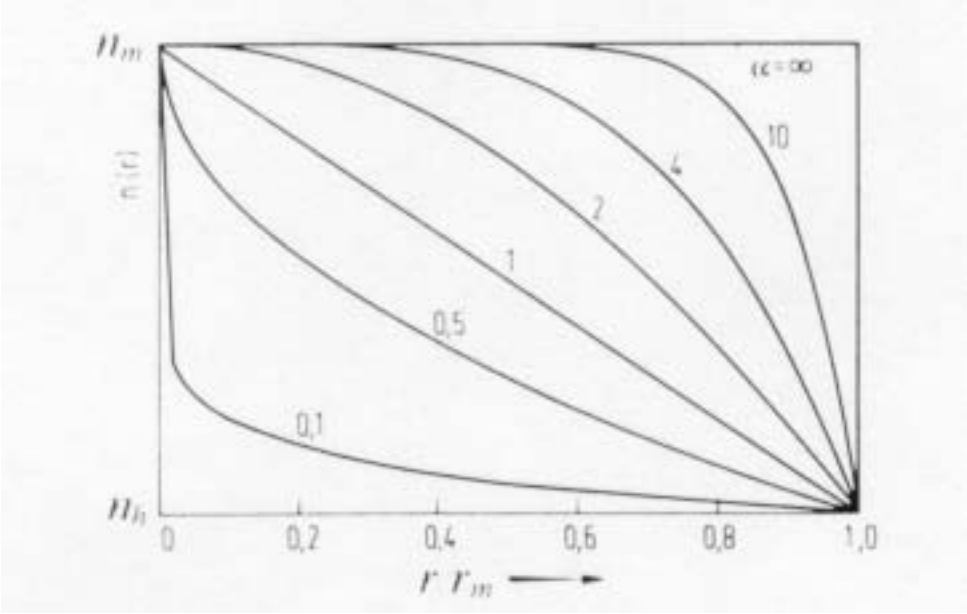


Figure 2.1.4.2. Power-law index profiles

The mathematical form of these profiles

$$n(r) = n_m\left[1 - 2\Delta\left(\frac{r}{r_m}\right)^{\alpha}\right]^{1/2} \qquad \text{ha} \quad r < r_m$$

$$n(r) = n_h \qquad \text{ha} \quad r > r_m \qquad (2.1.4.4)$$

where $\Delta = (n_m^2 - n_h^2)/(2n_m^2) \approx (n_m - n_h)/n_m$.

As can be proved the number of all of the propagating mode is

$$M = \frac{\alpha}{\alpha+2}\frac{1}{2}k_o^2 r_m^2 (n_m^2 - n_h^2) = \frac{\alpha}{\alpha+2}\frac{1}{2}V^2 \qquad (2.1.4.5)$$

where in the parameter *V* the maximum of the core index is to be written instead of the previous constant.

As the analysis pointed out in the case of $\alpha \approx 2(1-\Delta)$ extremely low intermodal dispersion can be achived.

2.1.4.2. Dispersion

The intermodal dispersion discussed above can be characterized by the intermodal dispersion coefficient

$$D_{im} = \frac{\Delta\tau}{L} \quad \text{[ns/km]} \qquad (2.1.4.6)$$

where $\Delta\tau$ is the group delay difference between the slowest and fastest mode and *L* is the length of the cable. For SI fibre

$$D_{im} \approx \frac{n_m}{c}\Delta, \qquad (2.1.4.7)$$

(where c=3.10^8 m/s), on the other hand for the optimal ($\alpha \approx 2$) parabolic profile

$$D_{im} \approx \frac{1}{2}\frac{n_m}{c}\Delta^2, \qquad (2.1.4.8)$$

which is about three order of magnitude smaller then the dispersion coefficient of the SI fibre. The intermodal dispersion appears also in the case of a strictly monochromatic light source. On the other hand the dispersion which is the

consequence of the finite spectral width of the light source is called *chromatic* dispersion.

Using a light source of finite spectral width, the impulse can be regarded as carried by the components of different wavelength independently, every with its own group velocity. According to this the impulse is widened. The difference of the maximum and minimum group delay is

$$\Delta\tau \approx \left.\frac{\partial\tau}{\partial\lambda}\right|_{\lambda_{vn}} \Delta\lambda\,, \qquad (2.1.4.9)$$

where $\Delta\lambda$ is the spectral width of the light source and the derivative is to be calculated at the nominal wavelength λ_{vn} of the light source.

The group delay difference per unit cabel length and per unit spectral width is called *chromatic* dispersion coefficient

$$D_{kr} = \frac{\Delta\tau}{L\Delta\lambda} = \frac{1}{L}\left.\frac{\partial\tau}{\partial\lambda}\right|_{\lambda_{vn}} = \left(\frac{\partial}{\partial\lambda}\frac{\partial\beta}{\partial\omega}\right)_{\lambda_{vn}} \qquad \frac{\text{ps}}{\text{km.nm}} \text{ or } \frac{\text{ns}}{\text{km.nm}} \qquad (2.1.4.10)$$

There are two kinds of the chromatic dispersion: the *material* dispersion and the *waveguide* dispersion. As a consequence of the wavelength-dependence of the refractive index the group delay is wavelength dependent even in infinite homogeneous material. Substituting into (2.1.4.10) the $\beta = k_0 n$ dispersion relation of the plane wave and changing the derivative according to the wavelength, the material dispersion coefficient is obtained

$$D_m = \frac{1}{c}\frac{\partial}{\partial\lambda}\left[n - \lambda\frac{\partial n}{\partial\lambda}\right] \qquad (2.1.4.11)$$

Because the wave is guided, the phase constant is a nonlinear function of the wavelength even in the case of a wavelength-independent index. The waveguide dispersion coefficient D_g can be calculated [on the basis of (2.1.4.9)] only numerically, because the dispersion relation of the fundamental mode, $\beta(\omega)$ cannot be given in analytical form.

In the dispersion relation, containing the wavelength-dependent index, the index and the $\beta(\omega)$ relation of the guidance are present in a fundamentally involved

form. Therefore the following relation to the total dispersion D_T by no means is justified theoretically, but can be used in practice

$$D_T \approx D_m + D_g \qquad (2.1.4.12)$$

The curves marked by CHANG are results of approximate calculation.

(see Figure 2.1.4.3.). As D_m is a monotonically increasing, while D_g is a monotonically decreasing function of the wavelength, the total dipersion coefficient can be made zero in some transmission window, with a suitable geometry (dispersion tayloring).

As can be shown the waveguide dispersion coefficient contains the same Δ factor, as the intermodal dispersion coefficient [see (2.1.4.7-8)] . To achieving small dispersion small indexdifference is choosed, but in this case -according to (2.1.4.1)- small V value is arised at a given wavelength and core radius, so the field is less concentrated into the core. Therefore such fibres are called *weakly guiding* fibres.

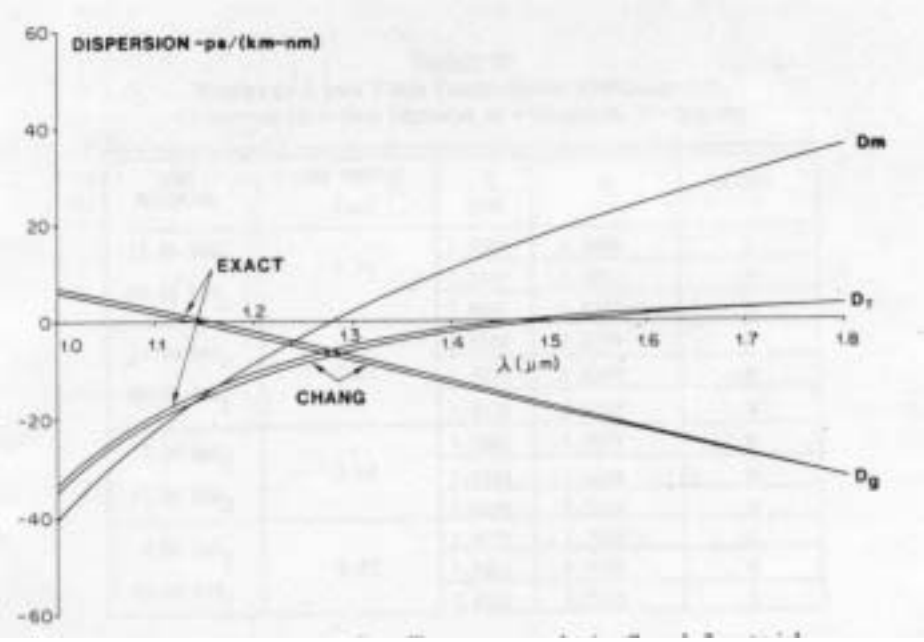


Figure 2.1.4.3. The material, the waveguide and the total dispersion coefficient for silicaglass.

2.1.4.3. Cable types

The fibres are coated by a polimer for the mechanical and chemical protection. The *primer coating* surrounds the fibre continously and its deposition is the organic part of the fabrication. The elastic position of the fibre in the cable is

assured by the *secondary coating.* Between the primer and secondary coating in a lot of case there is a *filling cushion.*

The arrangement in Figure 2.1.4.4. results a better coefficient of filling.

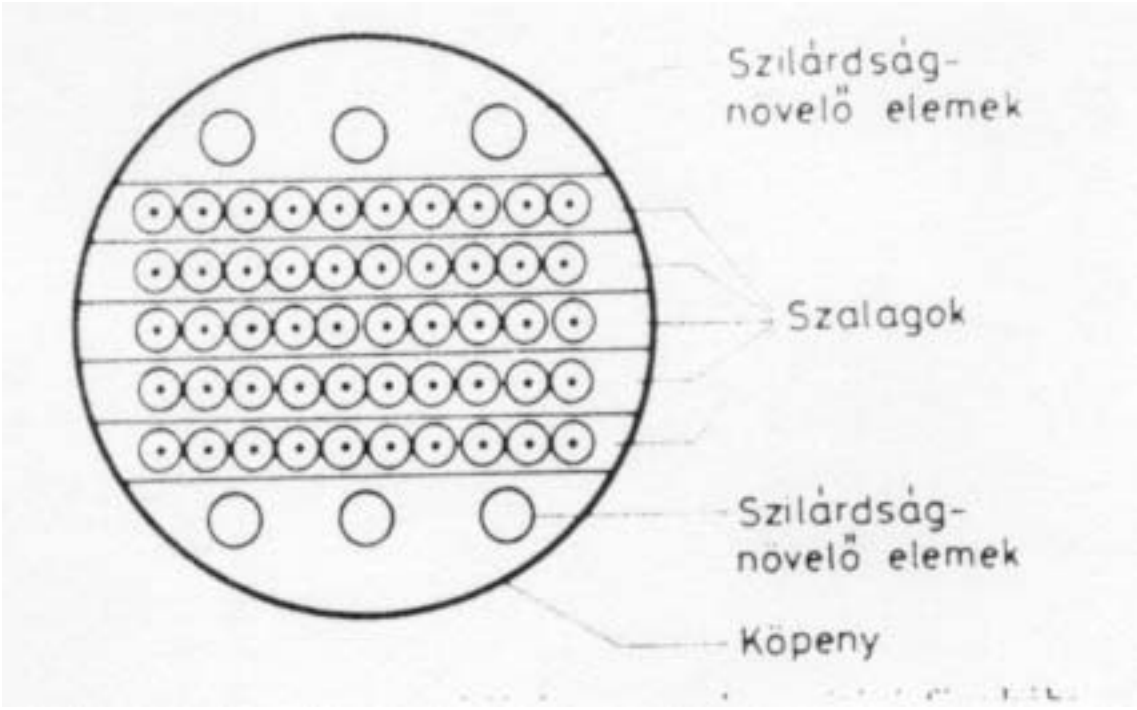


Figure 2.1.4.4. A large-capacity cable of ribbon structure

2.1.4.4. Connection of fibres

The fundamental object of the fibre connection technology is to make minimum the insertion loss. During the connection the following steps have to perform:

a. cleaning the fibre from the protection sheath
b. cutting of the fibre
c. accurate geometrical alingement of the fibres
d. fastening the fibres to each other
e. repairing the mechanical protection of the connection point

The *welded* connections means the fibre is melted and -during feeding- is fused. The melting is carried out mostly by electric arc, microflame or laser. In the case of *tubed* connections the positioning of the fibres is made by capillary. The fibres are introduced at the ends of the capillary. The fixing is made either by cement

(at the same time immersion material) introduced through the holes on the side wall of the capillary or by the heating and sintering of the capillary. At the *grooved* connections the positioning of the fibres is carried out by V-shaped grooves milled into plexi or metall. The advantage of the grooved connections, that they can be used as a group-connection method.

2.1.4.5. Laying, installation

From the point of view of the pulling-in the most important parameters of the cable are: the mass, the permitted largest maximum pulling force, the smallest bending radius and the ratio of the pulling force and the bending radius. The drawbacks of the *pulling in by manual force*, that it submits the cable to large mechanical load, the pulling-in forces cannot be measured and on acoount of the two kinds of frictional forces the starting pulling-in force is large. The *pulling-in by winch-machine* can be made by such a winch, which contains a pulling force measuring instrument, a pulling force limiter and a pulling velocity limiter. For avoid the tugging caused by the elasticity of the cabel the drum must be braked. On account of the damage of the cabels it is not recommended to pull in several cabels into a tube of large diameter. It is more correct to place in advance 3-4 inner tube of smaller diameter.

In the case of a cabel layed directly into the soil the digging up of the cable containing no metall is nearly impossible, therefore it is customary to lay a metall conductor upon the cabel.

References

[2.1.4.1] D. Marcuse: Light Transmission Optics. Van Nostrand, New York, 1972.
[2.1.4.2] S. Ungar: Fibre Optics. John Wiley, Chichester, 1990.

2.1.5. Line systems

Géza Paksy, author

István Frigyes dr., reviewer

The nodes of telecommunication networks are interconnected by wireline or wireless systems. Depending on the purpose of the network, the nominal length of these systems may be less than one kilometers or several thousands of kilometers. These systems have to transmit the electrical or optical signals carrying diverse services to the place of destination, with a limited delay time and free from bit errors, or jitter.

In course of the technical development, the first open wire or physical cable links have been followed by carrier systems and by the coaxial cable frequency division multiplex (FDM) analogue systems. In last third of the XX-th century the digital multichannel time divission system were preferred in new installating. Now the optical fibre digital systems carrying several ten thousands of telephone channels are widely used. Actually, the present research results of photonics indicate even higher transmission capacities by several orders of magnitude.

The most important function of line systems is to provide a transport service for the upper layers of the OSI system. This means that the line systems are handling the physical layer (with repeaters and propper interfaces) of the telecommunication networks, and are not concerned with the frame organization or packet structure of the signal carrying the service. This further items are not covering the following: internal content of the signal, signalling or addressing. Neither are investigated the integrity and qualilty of the circuits and packets providing the service.

The structure of line systems depends on their application area, and are matched to the properties of the transmission medium and the application circumstances. Their main parameers are the transmission capacity and the maximum repeater distance and the length of the ITU reference connection to which it can be fitled. Depending on these items, the following applications are the most important:

- intercontinental with systems extremely high (20000km) distances, mainly undersea cable systems and satellite transmission systems,
- national and international backbone high distance systems (max. 10000km)
- district network systems connecting settlements
- metropolitan network systems.

In the following, main parameters of only wireline (cable) networks will be considered. Wireless systems will be covered in Sec. 2.2.

General layout of line systems

The general layout of a point-to-point system is shown in Figure 2.1.5.1.

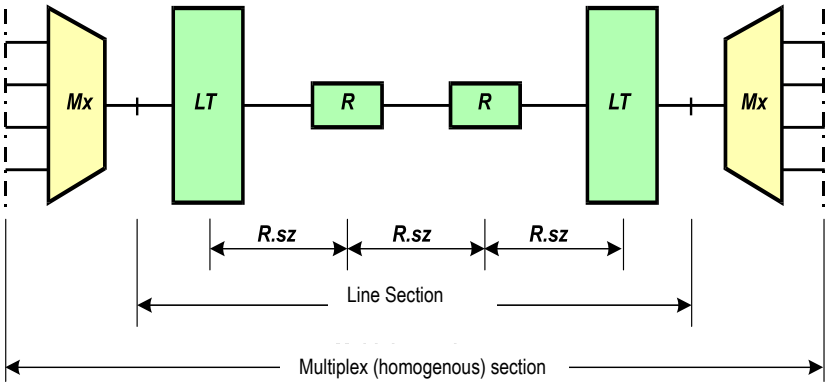


Figure 2.1.5.1. General layout of line systems

The LT functional blocks, providing the termination of the line systems, may be physically included in the multiplex equipment, but they usually constituted a separate unit, especially in out-of-date equipment, and were connected to the multiplex equipment over standard interfaces. In SDH systems, this separation is no longer present, and the line termination function is comprised in the interface ports [2.1.5.1].

Functional layout of line systems:

- Multiplex section: the section comprising equipment for multiplexing
- Line section: the LT equipment terminate the section comprising the transmission path (line), including also the line protection equipment.
- Regenerator section (denoted by R on the Figure): section between two regenerator stations
- Supplementary functions and equipment: data transmission channels for managing the network and for transmitting operational information. Providing auxiliary functions needed for the operation (remote supply, fault localization, monitoring transmission quality, optionally bit error correction).

Transport functions

- *Multiplexing*: the signals carrying the information can be multiplexed by time division, by frequency division, or by divison by any other orthogonal functions. These multiplexing procedures are detailed in Sec. 2.1.6.
- *Pulse regeneration:* this is a fundamental method of digital transmission, allowing the transmission of digital information, practically free of bit errors, over long-distance wireline systems.

Principle of regenerative repeaters

The application of regenerative repeaters allows the transmission of digital signals over theoretically arbitrary distances. The method has originally been developed for the first PCM systems using cables with copper conductors, but the principle is even today generally applied in other transmission systems too, independently of the transmission medium and the bit rate [2.1.5.2]. In the following, the principle of regenerative repeaters will be described.

Generation of the transmit signal

The electrical or optical representation of the binary form of the information to be transmitted, $\{b_i\}$, $-\infty < i < \infty$, can be expressed as followes:

$$S(t)= \sum_{i=-\infty}^{\infty} a_i \, s(t-iT),$$

where T stands for the pulse repetion time.

The transmit signal s(t) is usually an NRZ (non-return-to-zero) or RZ (return-to-zero) square pulse.

By further $\{b_i\} \rightarrow \{a_i\}$ encoding the $\{b_i\}$ binary information, the power density spectrum of the random transmit signal S(t) is usually shaped to obtain zero DC component stream. This can be achieved by coding in accordance with one of the DC-free line codes. The value of a_i can be 0,1 or +1, -1 in the case to two-level coding, or in the case of multi-level coding, N discrete equidistant values (e.g. +1, 0, -1), this is in most cases a quasi ternar stream, where +fand-1 has the same information content, but if +1,_1, sequence is hurted, has some meaning. E.g. the HDB-3, 4B3T, 8B110B or 2B1Q codes fall into this category.

The principle of regenerative repeaters is shown in Figure 2.1.5.2.

1. Reshaping

While the pulse train S(t) passes over a dispersive and noisy medium (copper or optical cable), the signal will be distorted and become noisy, depending on the applied medium. At the input of the regenerator, the receiver filter and amplifier has the function of shaping the signal to provide a form that is optimum for the decision making (see the Nyquist criteria), and amplifying it. In case of an optimal signal shape, the inter-symbol interference and noise will be minimal. The quality of the shaped signal is characterized by the “eye-diagram”.

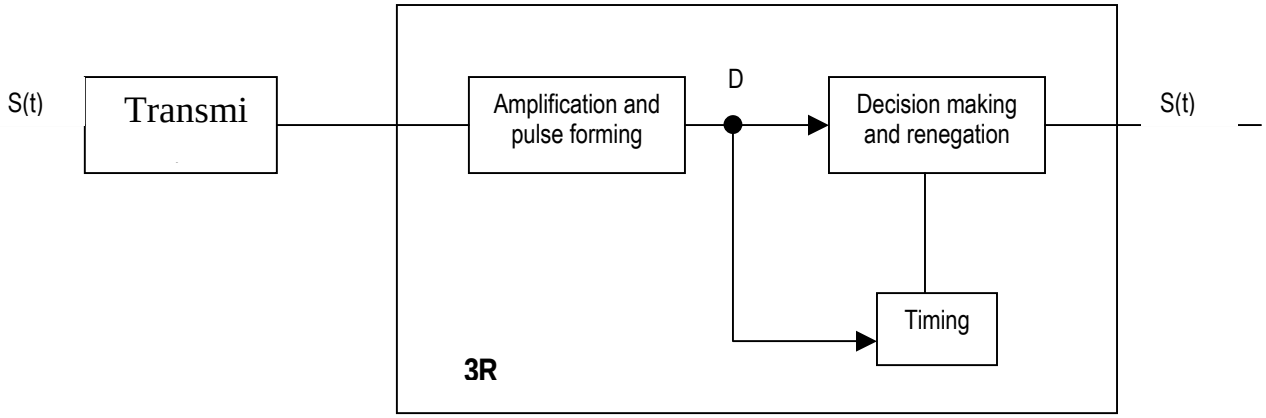


Figure 2.1.5.2. Principle of a regenerative repeater

2. Retiming

Normally, nonlinear signal processing is used to recover from the received pulse train the T clock signal having the original pulse repetition time. The recovered clock signal has two functions: partly, it serves to set the exact decision time instants, and on the other hand, it also restores exactly the width of the regenerated s(t) pulse. Practically, either SAW filters or PLL circuits are applied for restoring the clock signal. Imperfect clock signal restoration results in an additional noisy phase modulation of the restored clock signal called jitter. This has the effect of displacing the decision time instants, and possibly of false decision making. Over a long regenerator chain, these timing errors are accumulated, and at the receive side, may impair the quality of the transmitted signal, e.g. in the case of picture transmission.

3. Decision making and restoring the transmitted signal: Regeneration.

At point D of Figure 2.1.5.2, the values of $\{a_i\}$ have to be determined, the corresponding original S(t) waveform has to be conformed, i.e. the transmit pulse train S(t)has to be restored or regenerated at every time instant T, set by the timing signal. Decision is controlled by the timing signal, and the instantaneous value of the received signal has to be compared with the decision level depending on the number of transmitted levels. The decision may be correct or incorrect. The quality measure of the regeneration is the bit error rate, or more precisely, the $\{a_i\}$ line symbol error rate. In correctly operating regenerators in metallic cable systems, the error rate is less than 10^{-6}, and in optical systems, it is less than 10^{-12}. When maintaining long line sections, it is required to monitor continuously the error rate over the line. This is feasible either by controlling the internal regularity of the line coding or by continuous parity control of the pulse train. In high-capacity long-distance systems, error correction (FEC) is frequently used. In the STM-64 and STM-256 systems, the use of one of the BCH codes is proposed by ITU Recommendation G.709.

The above signal processing method is usually called 3R regeneratiron (Reshaping, Retiming, Regeneration). 3R regeneration is applied in a variety of metallic and fibre optic systems. In long-distance transmission, several intermediate

regeneration points may be needed. In symmetrical and coaxial cable systems applying copper conductors, the high attenuation calls for regeneration at every 2 to 4 kilometers, and this is why these equipments are installed in unattended underground containers with remote power supply. In PDH or SDH systems operating over monomode optical fibres, the regeneration distance is 40 to 80 km.

In systems with transmission speed in the Gbit/s range, 3R regeneration is extremely expensive so only partial regeneration is performed by omitting the retiming operation. Based on the above definitions, this is usually called 2R regeneration.

3R and 2R regeneration is possible only by electronics means (in year 2001). The realization of 3R regenerators by optical means in still in the research phase.

Optical amplifiers

Due to the attenuation of optical fibres, the distance between an optical transmitter and receiver, in the case of 2.5 Gbit/s transmission rate used at present, cannot be more than 100 to 150 km. In conventional systems using a single wavelength, 3R regeneration is applied, i.e. line repeaters are used.

In WDM systems, even 40 to 80 wavelengths are transmitted in parallel, so it would be extremely expensive to apply renegerators for each wavelength at all repeater station. Optical amplifiers are capable to amplify simultaneously all wavelengths. A single amplifier is sufficient to compensate the fibre attenuation introduced in the previous section.

The principle of optical amplification has already been conceived in the early stage of optical communication but actual application of this principle has taken place only in the mid-nineties [2.1.5.3]. There are several possibilities of relaizing optical amplification, such as using

- Erbium Doped Fiber Amplifiers (EDFA)
- Solid state optical amplifiers,
- Laser amplifiers,
- Optical amplifiers operating on the excited Raman scattering pricniple,
- Optical amplifiers operating on the excitingBrillouin scattering principle.

In present communication systems, only EDFA amplifiers operating in the 1520 to1560 nm range are applied, so in the following, only these will be investigated.

Erbium doped fibre amplifiers (EDFA)

Physical fundamentals

After doping the core of an optical fibre by a rare earth element, its atoms will be ionized when exciting them by a light ray of suitable wavelength $\lambda_{p,}$ and their free electrons will be shifted to a higher energy level. If these electrons return to the lower state, a light beam of wavelength λ_s will be radiated. The relation between the energy levels and the wavelengths will be given by the well-known equation

$\lambda=hc/\Delta E$

where λ is the wavelength of the radiated light beam, h is the Planck constant, and ΔE is the distance between energy levels.

The above relation calls for an Erbium doping agent in the wavelength range of 1520 to 1550 nm, and the irradiating light beam should have a wavelelngth of 980 or 1480 nm.

This is illustrated in more detail in Figure 2.1.5.3 depicting the electron energy levels of the lower state and the excited state. Responding to an incident light with a wavelength of 1480 nm, the electrons are shifted to an energy level of $^4I_{13/2}$, and subsequntly return to their original energy level while radiating a light beam of 1550 nm wavelength. The photons of the light having a power of P_{in} interact with the electrons being at the excited energy levels, and the power of the incident light of wavelength 1550 nm will be increased to P_{out} i.e. an amplification of $G = P_{out}/P_{in}$ will be obtained. An irradiation at 980 nm will result in a two-step enegy level shift, but the procedure will have the same result. The excited electrons return to their lower state by themselves while emitting light. This spontaneous emission appears as noise at the output of the amplifier. The photons appearing as a result of the spontaneous emission can further be amplified by the EDFA by spontaneous emission. This phenomenon is called Amplified Spntaneous Emission or ASE.

At the output of an amplifier of gain G, the ASE noise has a value of

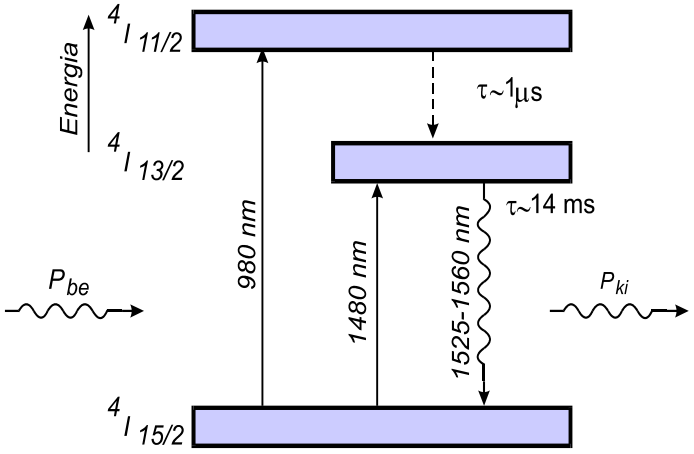
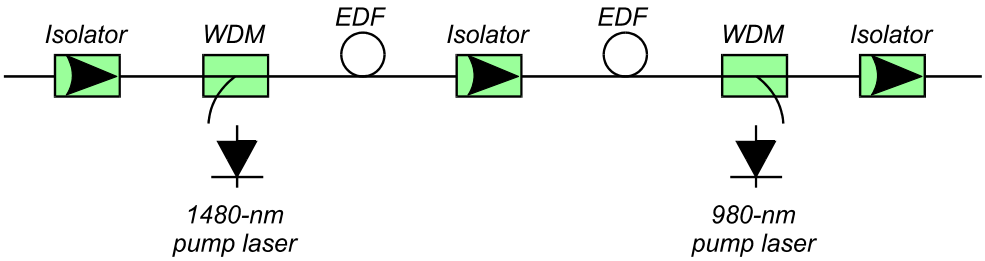


Figure 2.1.5.3



EDF - Erbium-doped fiber

WDM - Wavelength division multiplexer

Figure 2.1.5.4. Practical realization of an EDFA-type optical amplifiers

$S_{sp}= n_{sp} (G-1)hv$

Where n_{sp} is the spontaneous emission factor, having a minimal value of 1, and v is the light frequency.

The noise figure of the amplifier, based on the definition $F=(signal/noise)_{in}/(signal/noise)_{out,}$ can be calculated from the relation

$$F=2 n_{sp}(G-1)/G$$

Practical realization

The practical relization of the optical amplifier is shown in Figure 2.1.5.4, showing the block diagram of a two-stage optical amplifier. Here EDF denotes the doped fibre, and the isolators are applied to provide unilateral signal transmission. The exciting optical power is routed via the passive optical coupler to the erbium fibre of a few tens of meter, responsible for amplification, and coiled up on a spool.

Technical data of EDFA amplifiers

Output power: max. 15 to 23 dBm

Small signal amplification: 20-30 dB, at an exciting power of +20…+30dBm

Bandwidth: 20-30 nm between 3 dB points, in the wave band of 1525 to 1560 nm. The response of the amplification can be improved by an external optical filter inserted into the signal propagation path.

Noise: Main noise souce is the ASE noise, typical noise figure is F=4-5 dB.

Cascading optical amplifiers

In extended optical links, optical amplifiers have to be inserted into each section in order to compensate attenuation of the optical fibres. For achieving correct operation of the cascaded sections, the cumulative effect of noises generated by the optical amplifiers and the limiting effects of diverse dispersion distortions have to be taken into account.

The noises generated by Amplified Spontaneous Emission are propagated along the link and amplified by the amplifiers following the noise source. The accumulation of the spontaneous emission noise has two consequences. On the one hand, it will decrease the signal-to-noise ratio at the input of the receive amplifier, and on the other hand, it will introduce limitation in the amplifiers at the end of the link, thus decreasing their small-signal amplification.

Let us assume that a link comprising N amplifiers has a length of L km so that the distance beween amplifiers is s = L/N. Let us further assume that each amplifier has a gain of G, and this compensates the attenuation of the previous section. This means that $G= e^{\alpha s}$, and the attenuation of the optical fibre is α dB/km.

The over-all noise at the end of the link willl be

$$S_N = 2NS_{SP} = 2n_{sp}(G-1)N = 2\,n_{sp}(e^{\alpha s}-1)N$$

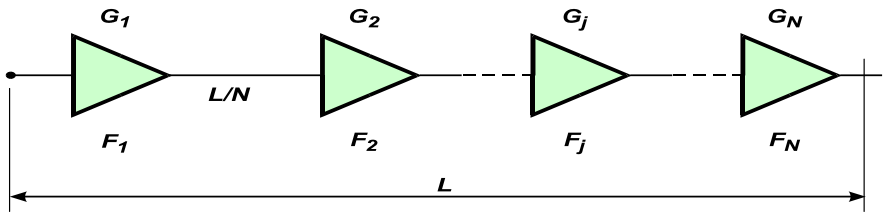


Figure 2.1.5.5. Cascade connection of optical amplifiers

The above relation shows that the noise is linearly increasing with the number of the cascaded amplifiers, and decreases with the length of the sections. Thus a viable solution would be to apply a high number of amplifiers with low section attenuation. However, the high number of amplifiers would increase the cost. As a compromise, distances between 50 and 100 km are usually applied.

Optical amplifiers can be applied not only as in-line amplifiers but also at the sending side in order to boost the optical power of the transmit laser, and to increase the receve level with a pre-amplifier.

Dispersion compensation

Chromatic dispersion is proportional with the length of the optical fibre and with the spectral width of the optical transmit signal.

The chromatic dispersion of the monomode optical fibre applied at present according to Rec. ITU G.652 has a positive value in the 1520 to 1600 mm range. This can be compensated by inserting into each section an optical fibre having opposite dispersion. According to Figure 2.5.1.4, these compensating fibres ar inserted between the stages of a two-stage optical amplifier. For this purpose, a special fibre

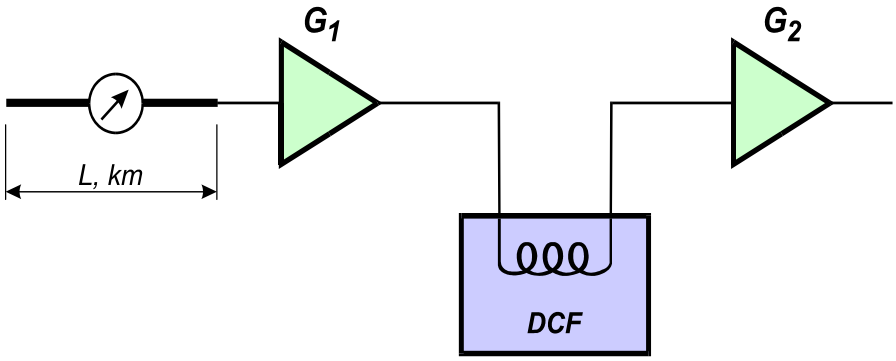


Figure 2.1.5.6. Compensation of dispersion in optical amplifiers

is used having a specific dispersion that is much higher than that of the compensating fibre. Thus a coiled-up fibre of appr. 1 km length is suitable to compensate for the dispersion of a complete transmission section.

This method is applicable up to bit rates of 10 Gbit/s. At higher bit rates, the polarization mode dispersion (PMD) has also to be compensated, in addition to the chromatic dispersion. Actually, this is the time difference between the modes of the optical waves propagated over the fibre. The PMD becomes critical when the time difference between the two modes, ΔT_{PMD}, becomes higher than one tenth of the transmitted signal bit period. An optical section of length L and specific dispersion D_{PMD} will introduce a time shift between the propagation modes given by

$$\Delta T_{PMD}=D_{PMD}\sqrt{L}$$

For standard optical fibres, the average value is D_{PMD} =0,4 ps/$\sqrt{L}$ km. At a transmission bit rate of 10 Gbit/s, one tenth of the bit period is 10 ps. From this, the highest distance that can be operated is calculated from the above restriction to be 625 km. However, for the fibre in the example, this distance at 40 Gbit/s bit rate is only 625/16=39 km so in this case, the PMD has to be compensated.

Basic types and restrictions of WDM equipment

The first generation of WDM systems was applicable only in long-haul point-to-point links, e.g. submarine intercontinental links. However, recent generations of WDM systems are suitable for realizing complex networks too, by approximately imitating the functions of SDH systems. To this end, the functions earlier applied for SDH digital systems had to be realized in the optical range. This principle has been utilized to realize the following optical equipment types [2.1.5.4].

Optically terminated multiplexer (OTM)

The OTM is suitable for receiving and transmitting N independent digital pulse trains over N independent wavelengths. The OTM is connected via protocol specific transponders to the client layers. Usual client signals are shown in the Table. In the case of protocol independent transmission, the bit error rate cannot be monitored.

Transponders	*Client signal*
SDH, ATM	*STM-16, STM 64*
GigabitEthernet	*1, 25 Gbit/s*
Fiber Channel	*1,062 Gbit/s*
Transparent protocol independent transmission	*0,1 – 2,5 Gbit/s*

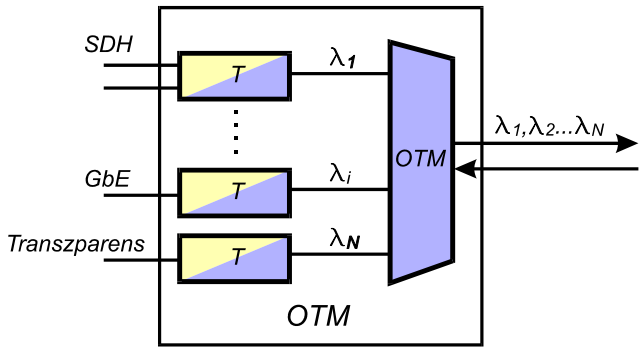


Figure 2.1.5.7. Functional representation of an optically terminated multiplexer OTM

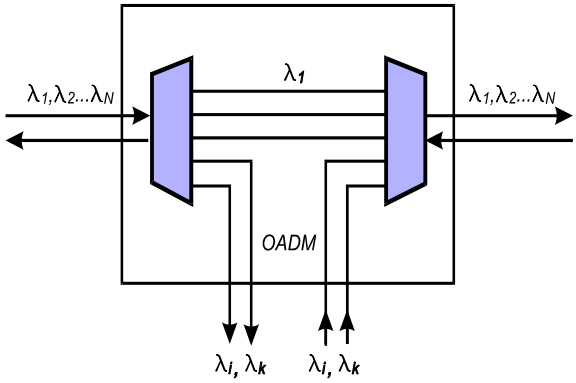


Figure 2.1.5.8.Functional representation of an optical add-drop multiplexer (OADM)

Optical add-drop multiplexers (OADM)

The OADM is suitable to branch n≤N wavelengths from N wavelengths, insert them into their place, and to transmit without changes the remaining (N-n) wavelengths. The operations are carried out in the optical domain so the OADM is transparent to the client signals given in the OTM description. The functional layout and physical realization of the OADM are shown in Figure 2.1.5.8.

Figure 2.1.5.9 shows the practical realization of the OADM. The Bragg grid filter rejects one wavelength (λ_2 in the example), this is reflected, and coupled out via the circulator. All other wavelengths are transmitted by the filter with low attenuation.

Utilization of the same wavelength by the in-coupled channel, and its insertion into the onward transmitted optical signal, is carried out by the secoond, right-side circulator. Coupling out several wavelengths requires the series connection of more filters. The insertion loss of these may be quite high so that an optical amplifier may also be necessary. The advantage of this solution is the simple layout, but it is

inflexible inasmuch that the coupled-out wavelength has to be permanently included in the equipment. It is expected that in the future, OADM types that can be configured dinamically for the complete wavelength range will be available.

OADM's will play a key part in the development of WDM self-healing rings.

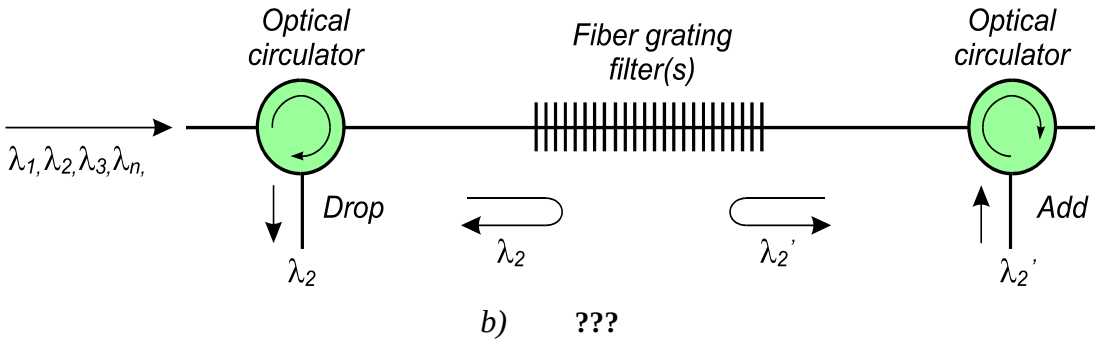


*Figure 2.1.5.9. Practical realization of an optical add-drop multiplexer (**OADM**)*

Optical cross-connect (OXC)

The controlled optical cross-connect arrangement OXC is suitable to re-arrange K optical fibres, each transmitting N wavelengths within an optical signal, into any of K optical fibres.

In principle, the OXC is capable of transposing the wavelength λ_n of an arbitrary fibre k_i into the wavelength λ_m of the fibre k_j, i.e of carrying out a wavelength conversion $\lambda_n \rightarrow \lambda_m$. According to the present state of art, this conversion can be carried out only in the electrical domain using O/E and E/O conversions. This means

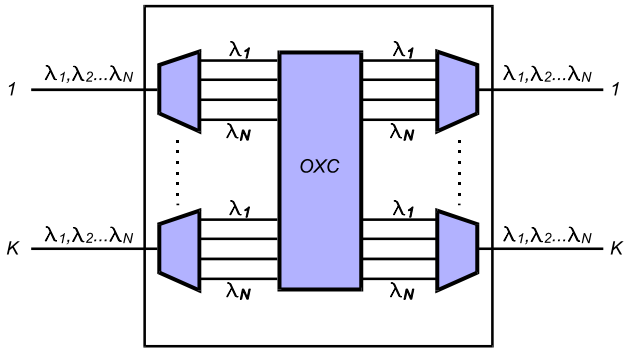


Figure 2.1.5.10.Functional schematic of an optical cross-connect (OXC)

that at present, OXC's are not transparent to optical signals.

Analysis of operating networks show that even in extended networks, only a limited number of conversions are necessary in a limited number of nodes. Several wavelength allocation algorithms are available to obtain a minimal length of paths, while the location and number of wavelength conversions are also limited.

In the year 2001, the manufacturing of optically transparent OXC equipment presented some problems. One of the problems was the switching of optical channels. At present, the use of micro electromechanical systems (MEMS) seems to be the most promising experiment [2.1.5.5].

References

[2.1.5.1] ITU-T Rec.G.901. General Considerations on Digital Sections and Digital Line Systems

[2.1.5.2] .G. Paksy: Transmission of digital signals, part of Lajkó-Lajtha: PCM in communication, Műszaki Könyvkiadó, Budapest, 1978

[2.1.5.3] Stern-Bala: Multiwavelength Optical Networks. Addison-Wesley, 2000.

[2.1.5.4] .D. Arató-R. Horváth-G. Paksy: WDM optical networks, PKI közlemények Vol.44. pp.73-111. Távközlési Kiadó 2000.

[2.1.5.5] Neukermans-Ramaswami: MEMS Technology for Optical Networking Applications, IEEE Communications Magazine, January 2001. pp. 62-69.

Translated by Tamás Sárkány dr.

2.1.6. Characteristics of the radio channel

István Frigyes dr., author

László Pap dr., reviewer

Introduction

Propagation characteristics of electromagnetic waves, their physical basis were discussed in Section 1.6. In this Section characteristics of wave propagation influencing transmission quality will be investigated. We'll mainly deal with the *microwave frequency range*, and with phenomena being important in the case of

digital transmission. These effects are, of course, harmful; as far as room is available, we'll also deal with countermeasures to these effects.

We have seen that if the distance between a transmit and a receive antenna is *D*, path loss is given as

$$a_0 = \frac{4\pi D^2}{G_a A_{eff,v}} = \frac{(D\lambda)^2}{A_{eff,a} A_{eff,v}} = \frac{(4\pi D)^2}{\lambda^2 G_a G_v} \qquad (2.5.1)$$

where subscripts *a* and *v* are used for transmitter and receiver, respectively,

G is antenna gain,

A_{eff} its effective surface,

λ is the wavelength corresponding to the carrier frequency

Magnitude of the path loss is really that, if the antennae are standing alone in the universe. Further path loss of a satellite link, operating with a frequency lower than about 10 GHz is not much different from that. A characteristic of the radio communication medium is, that the Earth environment influences in any other case it: signal-to-noise ratio is modified – usually decreased – linear distortion is caused. This type of phenomena is usually called *fading.*

A second characteristic of the radio channel is – contrasted to the wired one – that it is *open*. Consequently it is *not secure*, anybody can listen to it. Further – whether we want it or not – we receive transmission of radio links operating in our geographical neighborhood, which, very likely, will us disturb. This latter phenomenon is called *radio interference*.

A third characteristic is that if the channel is varying in time, due to any reason, frequencies are changed – both the carrier and its side-bands – and are also spread out. This phenomenon is called *Doppler effect*.

2.1.6.1. Fading phenomena

The following physical effects cause mainly fading:

Multipath propagation . Waves coming from the transmitter can arrive to the receiver over multiple paths if they are reflected, diffracted or scattered by objects of the surrounding. Of course these waves interfere with each other and thus they can

be strengthened or weakened. Multipath propagation can occur in any range of wavelengths.

Precipitation. Water molecules absorb electromagnetic energy. This phenomenon is characteristic at frequencies above about 10 GHz.

Gases. Gaseous molecules in the air absorb electromagnetic energy. This phenomenon is characteristic at frequencies above about 20 GHz.

Scintillation. In certain layers of the atmosphere – in some clouds, in surface layers – flow can become turbulent. As a consequence of that, refraction index of the air is randomly fluctuating causing fluctuation of the amplitude, the phase, the angle of incidence of the waves. This phenomenon is called scintillation.

Absorption caused by gases is more or less constant. All the others are changing randomly. Thus they can be described with statistical methods

2.1.6.2. Multipath propagation

Effects of multipath propagation are diverse in different systems. The most important ones are: mobile systems, fixed narrowband terrestrial systems and fixed wideband systems. We shall overview these after having summarized characteristics of time-variant linear systems.

2.1.6.3. Time-variant linear systems.

Their characteristics are needed for a precise description of multipath fading phenomena.

Received field strength is a vectorial sum of many rays; each of them contains the same signal with different time delays; thus the medium is *linear*. Delays and intensities of the rays as well as their number are randomly time variant.

Let the analytic signal of the signal radiated by the transmit antenna be

$$s(t) = u(t)e^{j\omega_c t} \qquad (2.5.2)$$

Arriving this over several paths the received analytic signal is

$$x(t) = \sum_n C_n(t)s[t - \tau_n(t)] \qquad (2.5.3)$$

where C_n and τ_n mean the loss and delay of the individual paths, respectively.

Corresponding complex envelope is

$$z(t) = \sum_n C_n(t)u[t - \tau_n(t)]\exp[-j\omega_c\tau_n(t)] \qquad (2.5.4)$$

Making mathematical treatment simpler assume that scatterers are continuously distributed. (2.5.4) can then be written in integral form

$$z(t) = \int_{-\infty}^{\infty} u(t-\tau)h(\tau,t)d\tau \qquad (2.5.5)$$

This is simply a convolution integral, however, impulse response h(.) is now time variant.

A time variant system or network can be described by its time variant impulse response. Other *system functions* are also available for this purpose; these apply other combinations of independent variables τ, t and frequency f, Doppler-shift ν. (Note that ν is sometimes called Doppler-frequency). Definition and application of these are summarized in equations (2.5.6)-(2.5.8)

$$z(t) = F^{-1}[U(f)T(f,t)] = \int_{-\infty}^{\infty} U(f)T(f,t)e^{2\pi jft}df \qquad (2.5.6)$$

$$Z(f) = \int_{-\infty}^{\infty} U(f-\nu)H(f-\nu,\nu)d\nu \qquad (2.5.7)$$

$$z(t) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} u(t-\tau)[S(\tau,\nu)]e^{2\pi j\nu t}d\nu d\tau \qquad (2.5.8)$$

Among these system functions there exist relationships based on Fourier-transform; appropriate variable pairs are f-τ, ν-t:

$$h(\tau,t) = F^{-1}{}_{(f)}\left[T(f,t)\right]; H(f,\nu) = F_{(t)}\left[T(f,t)\right]; S(\tau,\nu) = F^{-1}{}_{(f)}F_{(\tau)}\left[T(f,t)\right] \qquad (2.5.9)$$

Relationships are given above to time-variant transfer function T. F and F^{-1} mean Fourier-transform and its inverse, respectively; subscripts mark the appropriate transform-variable.

As already stated, channel is varying in a random manner; thus system functions can be modeled as stochastic processes depending on two independent

parameters. Very often what we know about these is not more than their correlation functions; it can be shown, that similar relationships exist among these as given in (2.5.9), but now we have *double-Fourier-transforms*.

We can assume in most cases that time varying transfer function *T(f,t)* is wide sense stationary both in frequency and in time. Thus its correlation function can be written as $R_T(\Delta f,\Delta t)$, i.e. it depends on frequency and time *differences* only, being independent both of starting frequency and of starting time. Such systems are called Wide-Sense Stationary, Uncorrelated Scatterers, abbreviated as WSSUS.

Coherence band B_C of a WSSUS channel is that frequency band in which

$$R_T(\Delta f = 0, \Delta t = 0) \approx R_T(\Delta f \le B_C, \Delta t = 0) \qquad (2.5.11)$$

(without specifying the degree of approximation). Similarly in its *coherence time* t_c

$$R_T(\Delta f = 0, \Delta t = 0) \approx R_T(\Delta f = 0, \Delta t \le t_C) \qquad (2.5.11)$$

Channel is wide band (or fading is flat) if spectral width of the transmitted signal $W < B_C$; otherwise it is *selective*. It is *slow* if duration of an elementary signal $T < t_C$; otherwise it is *fast*.

Correlation function of the time-varying impulse response can be written as

$$R_h(\tau, \tau'; \Delta t) = P_h(\tau; \Delta t)\delta(\tau' - \tau); P_h(\tau; \Delta t) = \mathbf{F}^{-1}{}_{(f)}\left[R_T(\Delta f; \Delta t)\right] \qquad (2.5.12)$$

P_h is the (inverse) Fourier transform of a correlation function; thus it is the *power density* of *h*. It specifies the time duration of the response to a narrow pulse, or to any other wideband signal. $P_h(\tau)$ is called *delay profile*; its rms value is the *delay spread*.

A *wide-band* channel attenuates all of the frequency components of the signal identically; such a channel causes thus mere attenuation of the signal. In a *selective channel* different components are attenuated by different magnitudes; so this channel causes linear distortion; we know that this results in intersymbol interference. In this case delay spread is large. (As: if the support of R_T is small, that of P_h is large and vice versa.)

Similarly, if *coherence time* is small, loss is varying within interval *T*, and the signal is distorted. With a similar arguing as above we can see that coherence time is small if Doppler shift is spread over a large band.

2.1.6.4. Propagation features in mobile systems

Effects of multipath fading are not the same in different environments. In present day telecommunications *urban* environments are the most important. Radio link is formed there between a traveling vehicle and a fixed base station. Buildings are usually high and consequently – in spite of having the base station antenna at high elevation – cannot be seen from the vehicle. Speed of the vehicles is usually high (not specifying more precisely what is meant under high speed). In *suburban* environments buildings are lower and built less densely; consequently a line of site between transmit and receive antenna is likelyexisting. A third type of environment is the *indoor*. This is the most hostile one; presence of the walls and movement of persons causes multipath propagation; further time variation is very slow, resulting in long duration of deep fades.

i. Doppler effect in this case is due to the relative movement of transmitter and receiver. If speed is v and the angle of incidence (relative to that of v) is γ, Doppler shift is

$$\nu = f\frac{\mathrm{v}}{c}\cos\gamma \qquad (2.5.13)$$

Rays are arriving from different directions and so Doppler-*spread* occurs. Its spectral density is given as

$$S(\nu) = \begin{cases} \dfrac{1}{2\nu_m}\dfrac{1}{\sqrt{1-\left(\nu/\nu_m\right)^2}}; & |\nu| \le \nu_m;\ \nu_m = f_c\mathrm{v}/c \\ 0; & |\nu| > \nu_m \end{cases} \qquad (2.5.14)$$

In (2.5.14) it is assumed, that that rays arrive in the horizontal plane and azimuth distribution is uniform. It is seen, that S(ν) is strictly band limited, and also that it becomes infinite at the band-edges.

ii. Statistics of the received signal – short term fading. In a short term received signal is WSSUS. If number of the rays is high, what can be assumed, *z(t)* can be

regarded as a Gaussian process (due to the central limit theorem). In urban environment its expected value is zero; average received power is σ^2. Absolute value *b* of *z(t)* is then of Rayleigh distribution, with probability density function

$$p(b)\frac{b}{\sigma^2}\exp\left[-b^2/2\sigma^2\right] \quad (2.5.14)$$

It can be shown that the phase of the received signal is of uniform distribution between 0 and 2π.

In suburban environment there is a *specular* ray. Components of the complex envelope are still gaussian with nonzero expected value. Absolute value is then of Rice distribution, with pdf

$$p(b) = \frac{b}{\sigma^2}\exp\left[-(b^2 + b_s^2)/2\sigma^2\right]I_0(b.b_s/\sigma^2) \quad (2.5.15)$$

Here b_s is magnitude of the specular component;

I_0 is the zero-order modified Bessel function of the firs kind;

σ^2 is the average power of the random signals (not containing that of the spacular component)

iii. Further variations of the received signal *Long-term fading* are mainly due to geographical conditions, variable shadowing, end of shadowing, etc. Its statistics are rather well described by lognormal distribution.

A third type of variation is the change of average received power due to change of transmit-receive distance. We know, that in free space field strength is inversely proportional to the distance of the antennas; in the presence of flat Earth to the distance squared. In the built-in environment, being of highest interest in the present chapter, variation is still inversely proportional to a power of the distance the exponent being dependent on the frequency and also on the type of buildings. As a rough estimate we can take the median of the field strength being inversely proportional to the fourth power of the distance.

2.1.6.5. Fixed terrestrial radio – narrow band transmission

In a fixed radio link visibility can be assumed. Consequently this medium is much more “quiet” than the mobile channel. Multipath propagation is still not excluded: characteristics of the atmosphere can be significantly different from the average one, although with low probability. This may result in anomalous spatial variation of the index of refraction of the air; as a consequence, a refracting atmospheric layer is formed and the refracted wave can interfere with the direct wave. This interference can result in a rather large attenuation. Temporal variation of this attenuation is slow consequently it doesn’t cause Doppler-effect; distortion isn’t caused neither, as the signal is narrow band. Empirical formulae exist for the attenuation.

According to experience complement cumulative distribution function of multipath-induced loss is

$$F(A) = \frac{6.10^{-7} xfD^3}{A}; A \geq 10 \qquad (2.5.16)$$

with *A* the fading loss, expressed in power ratio

f the frequency in GHz

D the distance in km

x an empirical constant depending on the climate and the; its magnitude at temperate climate and average soil is .25; at tropical climate and humid soil it is higher, at cold climate and rocky ground it is lower.

2.1.6.6. Fixed terrestrial radio – wide band transmission

Interference of rays is certainly a function of rapid variation of frequency. Thus in the case of wide band transmission loss may be non-constant even in the band of the signal, i.e. the fading may become selective. In principle system functions are again random processes, but these can be characterized in a simpler way as done in point *a*, i.e. with formulae containing random parameters.

Transfer function is again appropriate for describing the channel. If the spectral width of the signal is lower than 40 MHz, the so-called *simple three-ray model* can be applied. Its transfer function is

$$T(f) = a\left(1 - be^{2\pi j(f - f_0)\tau}\right) \qquad (2.5.17)$$

This formula describes a transfer function of periodic variation; loss is a cosinusoidal function of frequency. There are 4 parameters: *a* is the average of the transfer function; this multiplied by 1-*b* is its minimum; f_0 is the location of one minimum and $1/\tau$ is the distance of two neighboring minima.

Several different statistics are given in the literature for these parameters. In the so-called Rummler model

a is lognormal

the dB value of 1-*b* is exponential

τ is a constant (6.3 nsec, according to Rummler)

f_0 is two-level uniform.

2.1.6.7. Precipitation effects

Millimeter waves interact with any form of precipitation. Rain is of overwhelming significance. It has two effects: it causes loss and changes polarization states, decreasing thus cross-polar discrimination (XPD).

Rain induced loss depends mainly on rain-rate (mm/h). In the usual statistical approach the problem is divided into two parts: distribution function of the single-point rain rate is determined; and loss of a path is determined, conditioned on point rain rate. Several models are available for both steps. Concerning the first one, a world map is published by ITU-R, subdividing the Earth into a few climatic zones and giving the cumulative distribution functions valid for these zones. Of course, these distribution functions are not very precise; to get more precise data, measurements have to be made in the relevant area. Note, however, that these measurements – for yielding significant results – are rather lengthy and rather expensive.

For determination rain induced loss, method of ITU-R Recommendation 530-6 is described below.

The method is apt to determine fading loss of prescribed probability on a path. As input data length L of the path, $R_{.01}(1)$, the point rain rate averaged in one minute of probability .01%, further factor k and exponent α (being characteristic for the relevant frequency, geographical region and also to some extent the path itself). Then loss A_p occurring with probability p can be computed:

$$d_0 = 35\exp\left(-0{,}015R_{0,01}(1)\right)$$
$$r = \frac{1}{1+L/d_0} \qquad (2.5.18)$$
$$A_{0,01} = kR_{0,01}(1)^{\alpha} Lr$$
$$A_P = A_{0,01}.0{,}12P^{-(0{,}546+0{,}043\lg P)}$$

Among other forms of precipitation hail causes a large loss, possibly even larger than rain. However, very heavy rain and hail are similar phenomena, both occur mainly in summer thunderstorms. Thus treating separately is not justified. A few publications only are available in this topic.

Sleet is also important. A few publications report on its dominant role. However, a detailed theory or even significant amount of investigations is not available.

Among the indirect consequences of precipitation we remark that in winter months deep fades were experienced in periods of no precipitation at all. It turned out that these were consequences of reflections due to ice-layers on top of the snow. These can reflect millimeter waves causing similar multipath effects as overwater propagation.

Polarization-related phenomena form another class of precipitation effects. These can be of importance in spectrum-efficient systems called frequency-reuse systems, applying both polarizations for the transmission of two separate channels. XPD, being in non-rainy periods determined by the antenna, can drastically be decreased due to rain. This is based on two different physical mechanisms. One of these is polarization-dependent loss of raindrops due to their non-ideal spherical shapes. The other is change of the polarization state of the wave due to slant oval raindrops. There are several empirical formulae available on the statistics of XPD, however, as far as known by the author none of them claims to be valid above 35 GHz; therefore we omit their discussion here. We mention only that due to heavy rain the average XPD can be decreased to 15 dB and its minimum even to 9 dB (!).

2.1.6.8. Scintillaton

In certain layers of the atmosphere – in clouds, in surface layers – air-stream can become turbulent. As a result amplitude, phase and angle of incidence are randomly fluctuating, a phenomenon called scintillation.

Scintillation is characteristic in the upper microwave and in the millimeter wave band. Not dealing with its physics we summarize the effects influencing system performance. Probability of scintillation is much higher than that of heavy rain but the relevant fluctuation is much slighter; maximal attenuation is about 10 dB. Further: although this loss is frequency selective but coherence bandwidth is in the order of GHz. Consequently in present day systems its distorting effect is negligible

2.1.6.9. Atmospheric absorption and dispersion

If electromagnetic waves are propagating in media of low loss the effect of the medium can be characterized with a complex permittivity, imaginary part of which is significantly smaller than its real part. Loss due to the electrically neutral molecules of the air can be characterized like that and is usually negligible. Electrically asymmetrical molecules cause higher loss; it is reasonable to characterize them more precisely.

Of this type is the molecule of water vapor and also of several gases of lower density being of little significance. Water vapor has three resonance lines in the millimeter band, about 22.3 GHz, 183.3 GHz and 323.8 GHz; these are slightly depending on the temperature. Molecule of oxygen is paramagnetic, its magnetic moment causes similar resonance phenomena. It has an isolated resonance line at 118.74 GHz and a relatively broad absorption band at 50-to-70 GHz, composed of several resonance lines. A method to compute the attenuation due to gasous absorption is given in Recommendation 676-2 of ITU-R.

A few figures: attenuation above 20 GHz, depending on air moisture is nowhere lower than .05-.1 dB/km. About 60 GHz it is always higher than 10 dB/km.

Dispersion is a second potential consequence of molecular interaction. Attenuation and group delay of the medium is frequency dependent causing linear distortion. Dispersion is significant in the absorption band of 60 GHz, however

bandwidth of the medium is in the order of GHz; therefore the distortion is non-negligible only in the transmission of signals in the order of Gbit/sec. Present-day systems are of much lower bitrate.

2.1.6.10. Effects of fading

i. *Time-and-frequency flat fading* increases path loss decreases carrier-to-noise ratio. Its effect can be represented in two ways. We can regard carrier-to-noise ratio as a random variable; then the relationship of error probability valid in a Gaussian channel is regarded as a conditional probability and the total probability is computed. E.g. regarding PSK, error probability for a Gaussian channel is

$$P_E = 1/2\,\mathrm{erfc}\sqrt{\frac{E}{N_0}} \qquad (2.5.19)$$

On a fading channel this is conditional probability of error, conditioned on E/N_0. If the channel is of Rayleigh fading (received energy is then of exponential distribution) the total probability of error is

$$P_E = \frac{1}{2}\int_0^\infty \mathrm{erfc}\left[\sqrt{\frac{E}{N_0}}\right] p_{E/N_0}(E/N_0)\,d\left(\frac{E}{N_0}\right) = \frac{1}{2}\left(1-\sqrt{\frac{\bar{E}/N_0}{1+\bar{E}/N_0}}\right) \approx \frac{1}{4\bar{E}/N_0} \qquad (2.5.20)$$

Relationship is similar in the case of other modulations, inversely proportional to E/N_0.

According to the other concept we define an acceptable maximal error probability; if error probability gets higher we take the connection as interrupted. (In practice the operator disconnects really this link.) In this concept we have two parameters: P_E and the probability of outage, P_S. P_S. corresponds to a certain E_S/N_0. Thus e.g. for PSK the average error probability and the outage probability are

$$P_E = \frac{1}{2}\int_{E_S/N_0}^\infty \mathrm{erfc}\left[\sqrt{\frac{E}{N_0}}\right] p_{E/N_0}(E/N_0)\,d(E/N_0); P_S = \int_0^{E_S/N_0} p_{E/N_0}(E/N_0)\,d(E/N_0) \qquad (2.5.21)$$

ii. *Frequency selective fading* causes linear distortion and consequently intersymbol interference. This can also be regarded as a delay spread exceeding the duration of one symbol; then one symbol is spread into the time slot of the adjacent one(s).

iii. *Time selective, frequency flat* or *fast fading* is characterized by a transfer function non-constant during one symbol time. Signal is distorted in this case also:

$$z(t) = \int_{-\infty}^{\infty} U(f)T(f,t)e^{2\pi jft}dt = T(0,t)\int_{-\infty}^{\infty} U(f)e^{2\pi jft}dt = T(0,t)u(t) \qquad (2.5.22)$$

Although as seen this is also a sort of linear distortion, (note that an operation y=**O**x is called linear if for $x=c_1x_1 + c_2x_2$ $y=c_1\mathbf{O}x_1+ c_2\mathbf{O}x_2$.) but is different from the usual distortion. It is called therefore *multiplicative noise*.

iv If transfer function is *both time and frequency selective* it causes both usual linear distortion and multiplicative noise.

2.1.6.11. Communications over fading channels

i. A plausible countermeasure to flat fading is to increase power, i.e. to apply what is called a fade margin. But we see, either applying (2.5.20) or (2.5.21) that this increase is very large, even 40…60 dB. Thus fade margin can be regarded as a brute force method.

According to a different philosophy we apply a *diversity* method. In this we transmit the signals simultaneously over two (or more) loosely correlated channels. Then probability of both being in outage is much lower than that for one of them. Such diversity effect can be achieved by receiving the signals by two antennas, sufficiently far away from each other (space diversity); by applying two channels of different frequency (frequency diversity); or of orthogonal polarization (polarization diversity). If *L* loosely correlated channels are applied and signals are appropriately combined, the error probability is

$$P_E < \frac{1}{4\left(\overline{E}/N_0\right)^L} \qquad (2.5.23)$$

Combination is optimal, if the individual received signals are multiplied by the conjugate of the estimated transfer function and these products are summed. (In a simpler though inferior quality method instead of combining we simply select the larger signal; this latter method is called selection type diversity, while the former is maximal ratio combining.)

Two additions to diversity. In contrast to the above methods, in which diversity functioning is mainly based on operations in the receiver, recently transmit diversity is applied with success. In this the signal is coded in two (or *L*) different ways and transmitted by different antennae. Further: a special way of frequency diversity is the so-called RAKE system. In this spectrum of the signal is spread so that the individual signals of a multipath channel be distinguishable, yielding thus the diversity effect.

ii. Of course fade margin is not an effective countermeasure to frequency selective fading. Effective countermeasures are: adaptive equalization, compensating distortion of the channel; diversity with the same functioning as in the case of flat fading (i.e. *L* channels instead of one), although with basically different operating mechanism; multicarrier transmission, converting the signal to several parallel signals and transmitting these over channels of different frequencies (achieving so that the *individual* channels are narrow band and consequently selective fading becomes flat). We note tha synergistic effect of the simultaneous application of more than one countermeasure.

2.1.7. Terrestrial and satellite microwave links

Éva Gödör, Tamás Sárkány dr., authors

István Frigyes dr., reviewer

Introduction

Considering the transmission medium, transmission paths and systems can be divided into two basic groups: the guided wave group and the radiated wave group. The second one has two subgroups, the point-to-point (microwave) and the ponit-multipoint (UHF, VHF, short-wave-mobile and broadcast) systems.

The microwave systems can further be divided to fall into one of the following groups:

- terrestrial systems operating within the radio horizon (LOS),
- over-horizon (OH) systems.
  OH systems include the following types:

- systems based on tropospheric scattering
- stratosphere platform systems (HAP)
- systems utilising meteorite tails
- satellite systems

The above systems are predominantly operated in the microwave range. As a reminder, the applied frequency ranges are summarized in Table 1 (data are for information only).

The numbers in Table 1 serve not only for simple classification but also provide, as the casket of Pandora, a source for conflicts. All links and networks would like to possess the optimum spectral range corresponding to their physical parameters. This problem is solved by international agreements reached by ITU and WRC. In these agreements, the structure and operational parameters of a reference network are defined, such as the frequency range, the satellite orbit position, the radiated power etc.

Both in LOS and OH links, the transmission medium includes the terrestrial atmosphere so its effects have also to be taken into account: radio window position, absorption lines of water vapor and oxigen, effects of precipitation.

Terrestrial links

Since the first FDM-FM microwave link transmitting 300 telephone channels and a video signal in the C-band has been put into operation between New York and Boston in 1947, several LOS systems have been developed [2.1.7.1], [2.1.7.2]. Let us now summarize the system types and the quality requirements.

The main types of LOS digital radio-rely systems meeting to-day's telecommunication requirements are the following:

- classified according to the distance covered:

Designation of frequency range	L	S	C	X	K_u	K
f [GHz]	1-2	2-4	4-8	8-12	12-18	18-27

Designation of frequency range	K_a	Q	V	U (USA)	W
f [GHz]	27-40	33-50	50-75	40-60	75-110

Table 2.1.7.1

- long distance system (backbone network);
- medium distance system (district network);
- short distance system (local network);

- classified according to the transmission capacity:
  - low capacity system (up to 10Mb/s);
  - medium capacity system (between 10 Mb/s and 100 Mb/s);
  - high capacity system (above 100 Mb/s);
- classified according to the network topology:
  - point-to-point links (P-P);
  - point-to-muiltipoint links (P-MP);
  - ring-type links;
  - mesh links.

  Note: in microwave systems, ring-type links and mesh links are rarely applied.

Another type of frequently used classification is the wideband/narrowband designation. However, this is rather ambiguous as it only refers to the transmission capacity. Note that the bandwidth of a modulated signal depends not only on the modulation signal properties but also on the type of modulation.

Up to the end of the mid eighties, the basic recommendation for both wireline and microwave links was the Rec. ITU-T G.821, applicable fo the 27 500 km long Hypothetical Reference Digital Link (HRDX). This reference path could be divided into two sections of local quality (10 km), two sections of medium quality (1250 km) and a single section of high quality (25 000 km)

This latter could further be divided into ten HRDP (2500 km) sections, each of them comprising nine HRDS (280 km) sections. Transmission quality was characterized by diverse type of bit error rates of the 64 kb/s transmission path, but these requirements could not be directly applied to the higher levels of the multiplexedbit streams. The quality of higher bit rate digital transmission paths is characterized by data derived from the number of errored blocks. This is the principle of Rec. G.826 adopted in 1993 and of Rec. G.828 adopted in 2000, respectively [2.1.7.3].

The requirements of these recent Recommendations are also applicable to a hyphotetical reference path (HRP) of 27500 km, but this system includes both distance-independent and distance-dependent requirements.

As an example, consider the transmission of an STM-1 signal over a microwave SDH system with a reference connection of 2500 km, in a moderate zone, over a flat and hilly terrain. To solve this task, the type of microwave link required by the task is selected in course of the network design (high capacity backbone network, P-P), and subsequently, the radio channel statistical properties applicable to the geographical area are determined.

One of the possible solutions for the task in this example is a 128-state QAM microwave equipment operating in the 6 GHz frequency range.

Over-horizon links

For links within the radio horizon, the antennas should be placed to provide direct line-of-sight transmission. Practically, this means that the eqiuipment should be installed on towers or buildings of a height, which fullfil the line-of-sight requirement and minimalize the probability of the fading due to multipath propagation.

For over-horizon links, a virtual tower hight is defined either by the propagation mechanism itself (scattering capability of the troposphere, reflection capability of the meteorite tail), or by the physical placement of the equipment (stratosphere platform, satellite platform).

System based on tropospheric scattering

The troposphere is the lowest 10 km part of the atmosphere. The inhomogenities within the common beam volume of the transmit and receive antenna act as sources scattering the electromagnetic waves into all directions of space. Thus components will be generated establishing coupling between transmitter and receiver. Tropospheric links are based on this weak coupling mechanism [2.1.7.4]. The first troposcatter link has been put into operation in 1953.

It follows from the geometry of the link that distances of up to 400 to 500 km can be spanned. This is why troposcatter links have been applied at places that were hardly accessible such as links connecting oil well islands with the mainland, links

within deserts and jungles, or links for military communication (before the large-scale use of inexpensive satellite links). These links are operating in the 350 MHz to 6 GHz range within the bands allocated by the ITU. The parameters of the radio channel require extremely high-gain antennas and transmitting powers, further the application of multiple diversity systems such as frequency and space diversity, angle and frequency diversity. The capacity of the transmitted information is limited by propagation time differences over the scattering volume. Typical values are the transmission of 120 to 240 voice channels in FDM-FM systems, or 8 to 12 Mb/s bit rates in FSK or PSK systems.

Stratosphere platform (HAP)

Equipment established in the stratosphere, i.e. installed on airhips, airplanes, helicopters or their combination, is collectively designated as High Altitude Platform (HAP) equipment. Links comprising this equipment will be part of the global telecommunication infrastructure, in co-operation with either wideband satellite systems or terrestrial WDMA optical networks or UMTS systems [2.1.7.5]. Frequency ranges for HAP systems have already been allocated by WRC-97, and later WRC-2000: the vicinity of 48 GHz, 18 to 32 GHz, 1885 MHz to 2170 MHz, and for interconnection of HAP's the adjacent IR range (800 to 1600 nm). At present, development is in progress in the fields of the required devices, circuits subsystems (tehnology of laser links between instabil terminal sites, phase controlled antenna systems, high efficiency solar cells and storage batteries). In addition, propagation measurements are in progress, especially for investigating propagation within buildings, further the screening effects due to terrain objects because of the low elevation angle.

Platforms are ready as the suitable devices had only to be selected from the available stock. At the Paris Air Show in 1999, a plane developed for HAP, designated PROTEUS, has been shown. Stratosphere platforms are suitable for meeting telecommunication requirements of habitations dealing with either heavy or light traffic. An example for meeting the former requirement is the point-to-point HAP system utilizing platforms located over high-seas navigation routes, thus providing backbone links. In the planned cellular systems, at least fifty electronically controlled spotbeams are applied, and the cell structure corresponds eiher to the conventional

terrestrial structure, or – according to another concept – will comprise concentric circular rings. There will be communication between the platforms themselves and also between any platform and the LEO, MEO and GEO satellite systems. However, this latter solution would significantly increase the path-loss and the propagation time, although the most important advantage of the HAP system, the short propagation time due to the short distance of appr. 20 km, would further on allow the use of protocols applied in terrestrial systems. The first HAP system was put into operation in 2002: this is the Sky Station system, comprising 250 equipments on airships filled with helium, operating in the 48 GHz band with a capacity of 2 Mb/s.

System utilizing meteorite burst

Meteorites entering the atmosphere are glowing and producing an ionized band lasting from a few tenth of a second up to a few seconds at a hight of 80 to 120 km. Depending on the density of the charged particles, this burst will either scatter or reflect electromagnetic waves in the 40 to 100 MHz range, allowing communication to be established over distances of 400 to 2000 km [2.1.7.4]. This phenomenon was discovered in 1935, and later on, between 1950 and 1975, intensive development work has been carried out in order to determine the properties of these radio channels. The first meteor-burst system started operation in Canada, in the year 1953, and in 1967, the ANOTEL (Snowpack Telemetry) system in Alaska was put into operation. This system was used in an uninhabited area for transmitting sensing data from 500 remote stations to a central station.

The period over which communication is established is determined by the life time of the ionized band. This is taken into account by storing the data and transmitting them in burst form, so this kind of transmission is called meteor-burst transmission. This is an inexpensive system, sometimes called “the satellite of poor people”, and is suitable to transmit data of a few kb/s. The Alaska system has been further developed, and in the technical literature, more civilian and military systems have been reported. One of these is the BLOSSOM system (Beyond Line-of-Sight Signaling Over Meteors) put into operation in 1986, operated in the 37 to 72 MHz band with FSK modulation and 2.4 kb/s capacity. Another tropo scatter system is operating in Egypt since 1995, transmitting water-level data of the Nile that are

needed for watering, and one more station is installed at the top of the volcano Mount St. Helens, collecting data received from stations monitoring snow thickness .

Satellite links

Devices installed on satellites are used for diverse purposes such as radio astronomy, radar astronomy, meteorology, remote sensing, navigation, telecommunication. Satellite telecommunication systems can be classified into three grouops: FSS (Fixed Satellite Service), MSS (Mobile Satellite Service), DBS (Direct Broadcast Service). Key words of satellite communication are: type of orbit, free-space attenuation, propagation time, frequency, polarization, type of beam, global coverage, multiple acces, cost.

Satellite orbits may have a circular or elliptical shape. There are several kinds of circuilar orbits: LEO = Low Earth Orbit, 400 – 1 000 km, MEO = Medium Earth Orbit, 5 000 – 13 000 km and GEO = Geostationary Orbit, i.e. in the equatorial plane at a height of 35 785 km.

Parameters, advantages and disadvantages depending on the orbit height:

- Satellite-Earth propagation time: 238 ms for GEO, 66 ms for MEO, 6 ms for LEO
- Number of satellites required for global coverage: three are sufficient for GEO (classical Clarke orbit), 10 to 15 for MEO, 40 to 60 for LEO
- Free-space attenuation follows the (distance/wavelength)^2 law, its value for GEO in the C-band being 200 dB.
- The Doppler frequency shift is proportional with the radial velocity between the satellite and the tererstrial station. For instance, f_d = 20 Hz for GEO for a distance shift of .001 km/s at 6 GHz, but for LEO, it may be as high as several hundreds of kHz.
- The cost of launching the satellite is highest for GEO, it equals with a great number of LEO satellites to be launched.

Typical frequency bands for communication satellites: 6 GHz uplink and 4 GHz downlink, in short 6/4 GHz (C-band), 14/2 GHz (K_u band) **??** , 30/20 GHz (Ka band), 50/40 GHz (V band).

The great variety of operating and planned satellite systems is partly due to the diverse application fields envisaged for a single system. Even a single item, e.g. personal communication, can be realized with several system philosophies. For

instance, on-board signal processing is used by the Iridium system while this is done by the Globalstar system at the ground stations.

The attributes in the system designations such as „little, big, mega, giga" refer to the data transmission rate. "Little" systems cannot be used for voice transmission, their typical task being credit card identification, container tracking, signalling of vehicle position, collecting data of utility meters (gas, electricity, water) and remote sensing data, and communication in emergency situations. "Big" systems can be used for voice and low-speed data transmission, further for paging. The "mega" systems are intended for realizing "Internet via satellite" and "Internet in the sky" [2.1.7.6], [2.1.7.7]. LEO-systems such as TELEDESIC, SKYBRIDGE and GEO systems such as ASTROLINK, ORION F-2, F-8, F-9, CYBERSTAR, WILDBLUE-1,2 are planned to realize these programmes. These planned wideband satellite systems, to be operated in the K_u band or even in the K_a band, will possess onboard signal processing, electronically controlled antenna beams and inter-satellite links. Suitably chosen modulation methods and error correcting coding will help these satellite channels to provide fibre-like service quality.

The INTELSAT-901 satellite, the first member of the new series launched June 9, 2001, has 22 high-power transponders operating in the K_u band, calculated in units of equivalent 36 MHz transponders. Taking this into account shows that the satellite wideband access will become a reality of the near future.

References

[2.1.7.1] I. FRIGYES: Communication Systems (in Hungarian), Műegyetemi Kiadó, 1998

[2.1.7.2] M.DREIS: Results of WRC-200 in relation to the Fixed Service,
Proceedings of Seventh European Conference on Fixed Radio Systems and Networks (ECRR 2000), Sept. 12-15. 2000 Dresden, Germany, pp. 9-14

[2.1.7.3] V.M. MINKIN: Impact of new Recommendation ITU-T G.828 on Design of Digital Radio-Relay Links
Proceedings of ECRR 2000, Sept. 12-15. 2000 Dresden, Germany, pp. 287-291

[2.1.7.4] Roger L. FREEMAN: Reference Manual for Telecommunications Engineering, Second Edition
J. Wiley, 1994

[2.1.7.5] Y.C. FOO, W.L. LIM, B.G. EVANS: Performance of High Altitude Platform Station (HAPS) CDMA System, Proceedings of 19th International Communications Satellite Systems Conference and Exhibition, 17-20 April 2001 Toulouse, France, Vol. 3 pp. 905-915

[2.1.7.6] Peter J. BROWN: K_a-Band Services, Via Satellite, Febr. 2001 pp. 18-28

[2.1.7.7] SHINGO OHMORI: Mobile Satellite Communications, Artech House, 1998

List of abbreviations in Sec. 2.1.7

ATM = Asynchronous Transfer Mode

BLOSSOM = Beyond Line-of-Sight Signaling over Meteors

DBS = Direct Broadcast Service

EB = Errored Block

FDM = Frequency Division Multiplexing

FEC = Forward Error Correction

FM = Frequency Modulation

FSK = Frequency Shift Keying

FSS = Fixed Satellite Service

GEO = Geostationary Orbit

HAP = High Altitude Platform

HRDP = Hypothetical Reference Digital Path

HRDS = Hypothetical Reference Digital Section

HRDX = Hypothetical Reference Digital Connection

IR = Infra Red

ITU = International Telecommunications Union

LEO = Low Earth Orbit

LOS = Line-of-Sight

MEO = Medium Earth Orbit

MSS = Mobile Satellite Service

OH = Over-the-Horizon

P - MP = Point-to-Multipoint

P - P = Point-to-Point

PSK = Phase Shift Keying

QAM = Quadrature Amplitude Modulation

QAM = Quadrature Amplitude Modulation

QPSK = Quarternary PSK

SDH = Synchronous Digital Hierarchy

SNOTEL = Snowpack Telemetry

STM-1 = Synchronous Transmission Module

UMTS = Universal Mobile Telecommunications Systems

WDMA = Wavelength Division Multiple Access

WRC = World Radiocommunication Conference

Translated by Tamás Sárkány dr.

2.2. Transmission devices and methods

In line with the previous treatment, this Section presents transmission devices and methods of communication systems in two parts covering wireless and wirebound systems.

Sec. 2.2.1 surveys the modulation and multiplex system such as PDH, SDH and WDM. Next, in Sec. 2.2.2, solutions intended to improve the exploitation of subscriber access networks are outlined: ISDN, basic, primary digital multiplexing, xDSL (ADSL, HDSL, ADSL, PON). These abbreviations are explained in the introduction to these Sections. Elements of cable TV networks are separately dealt with in Sec. 2.2.3 while Sec. 2.2.4 covers the wideband (34 to 2000 MHz) transmission methods and the fibre distribution networks (FTTH).

Sec. 2.2.5 starts by explaining the basic properties of wireless cellular systems (clusters, frequency re-use, interferences, efficiency parameters, capacity), and analysing other factors affecting system parameters such as cell sectorization, adaptive antennas, power control etc.

This is followed in Sec. 2.2.6 by presenting the specific wireless interface elements of satellite and terrestrial mobile systems. In this part, cordless phones (CT1, CT2, DECT), pagers (ERMES), trunked mobile systems (TETRA, TETRAPOL), further solutions applied in cellular mobile phone systems (first generation – NMT, second generation – GSM, IS-95), third generation – UMTS, WCDMA.

A separate Section 2.2.7 presents the specific elements of wireless office systems, covering the elements of wireless PBXs, WILL, WLAN (IEEE 802 11, HYPERLAN 1,2) and Bluetooth.

Sec. 2.2.8 presents systems at the borderline of informatics and telecommunications (mobile Internet, mobile IP, mobile software concepts, mobile agent, software radio). Sec. 2.2.9 is the concluding part of Sec. 2.2, presenting the transmission methods of terrestrial and satellite broadcast systems.

Sándor Imre dr. Editor of the Chapter

2.2.1. Digital PDH and SDH hierarchy

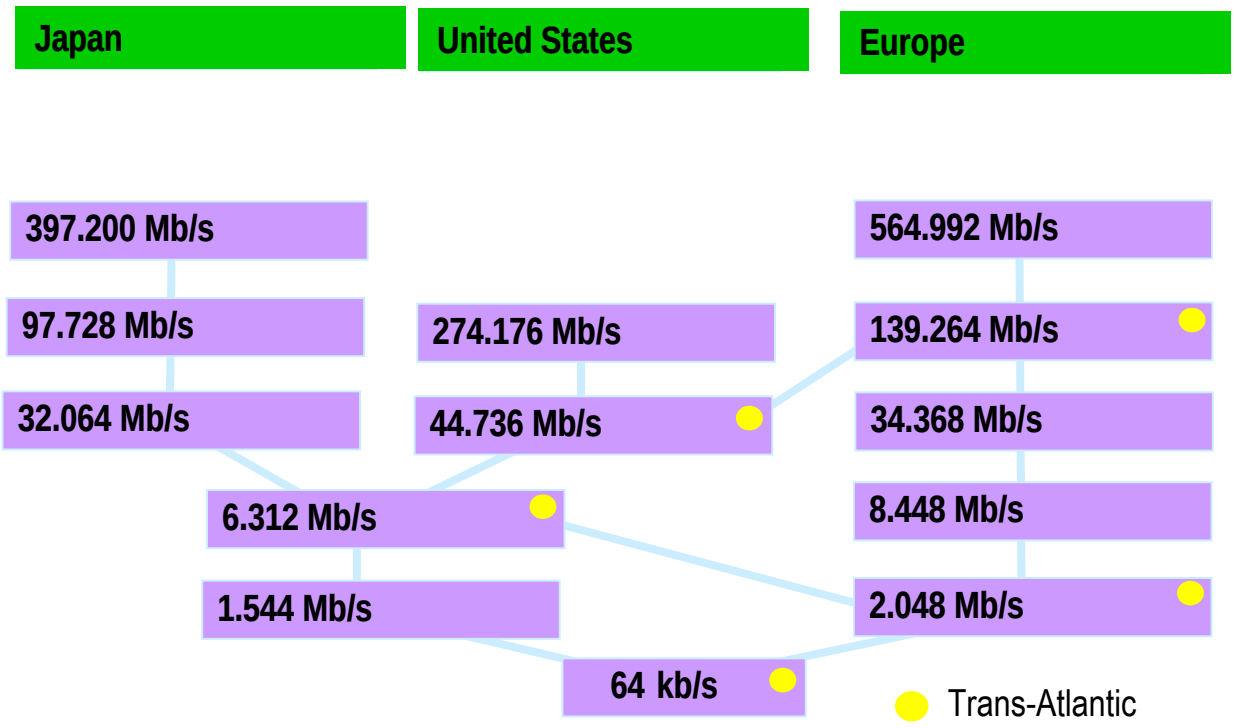
Tibor Cinkler dr., author

Géza Paksy, Péter Jeszenői, reviewer

PDH networks have been developed for multiplexing and transporting many telephone lines spanning the same path. However, as shown by the Figure below, a variety of multiplexing hierarchies have been adopted. As soon as communication between continents became viable, a solution for intervconnecting these system has been searched for, and also to find a system yielding higher reliability and flexibility, together with excellent data transmission possibility. These requirements paved the way for the SONET and SDH systems which will be outlined in the folowing.

2.2.1.1. PCM –(Pulse Code Modulation)

The analogue voice signal within the 300 Hz to 3400 Hz frequency band is sampled with 8 kHz according to the Nyquist-Shannon theorem (with a little allowance to take into account the imperfection of filters), and quantization with a compander is applied. The applied compressing functions used in Europe and



overseas are different.

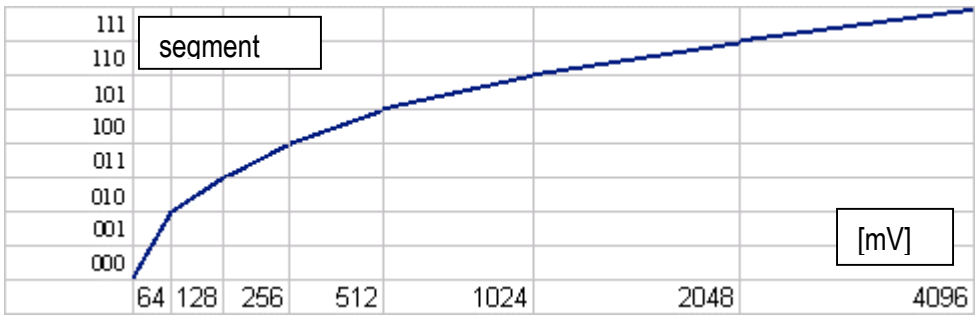
In Europe, the A-law is applied, the compressin function being

$$y=\begin{cases}\dfrac{Ax}{1+\ln A}, \text{ha} |x|<1\\ \dfrac{1+\ln Ax}{1+\ln A}, \text{ha } |x|\geq 1\end{cases}$$ A=87,6

In the United States and Japan, the *μ-law* is applied:

$y=\dfrac{\ln(1+\mu x)}{\ln(1+\mu)}$ where μ=256.

In practice, the broken-line approximation of this nearly logarithmic quantizing characteristic is used, allowing the direct implementation of the compressing and expanding (companding) functions, further the encoding and decoding functions. Companding has the advantage to allow the implementation of a specific signal-to-noise ratio with an 8-bit code instead of a 12-bit code that would be required without companding, in the same dynamic range. The characteristic below shows the broken-line approximation of the A-law.



The coding functions of an 8-bit code are the following: the first bit stands for the polarity (positive: 0, negative: 1). The following 3 bits designate the segment, and the fourth bit is for the linear characterisation of the signal value (see the Figure).

The absolute value of the signal falls between 0 and 4096, the boundaries of the 8 segments are at 32 mV, 64 mV, 128 mV, 512 mV, 1024 mV, 2048 mV and 4096 mV. Within any segment, linear quantization is applied but the first two segments (0 to 32 and 32 to 64 mV) have identical tangents. This means that the "resolution" of this code is 2 mV both in the 0 to 32 and 32 to 64 interval, and is deteroriating in each further segment (the "step" is duplicated): 4 mV, 8 mV, while in the last step it is 128 mV.

Example: encoding sample 1970. This is a sample with positive sign, falling into the segment 1028-2048. (1970-1028)/64=14.72, corresponding to the 15th linear interval. This means that the corresponding code will be 1110. (Because the code of the first interval is 0000, and that of the 16th interval is 1111.)

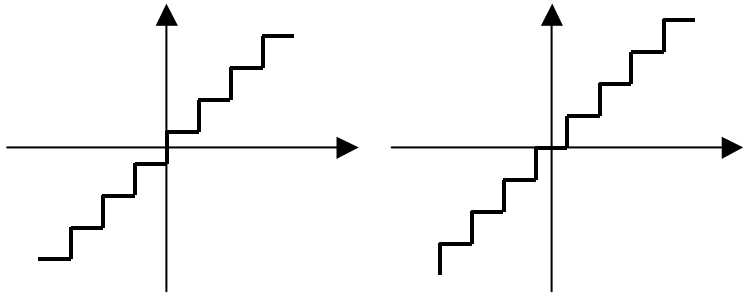
	Polarity	Segment			Linear encoding within the segment			
1970=	0	1	1	0	1	1	1	0

The behaviour near zero differs in the A law and the μ law. At zero, the midriser pertaining to the A-law exhibits a step while the midstep pertaining to the μ-law has a horizontal section, so the small noise around zero does not introduce fluctuations. The result of the encoding is thus a PCM (Pulse Code Modulation) signal at a rate of 8 kHz x 8 bit = 64 kb/s.

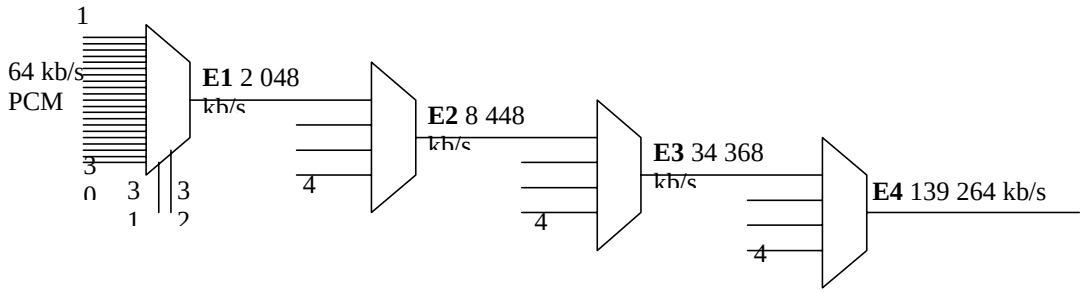
2.2.1.2. PDH - Plesyochronous Digital Hierarchy

The following multiplexing methods are applied in PDH telephone networks.

- Plesiochronous (nearly synchronous). The clock signals of individual network tools (multiplexers) are identical only within specified tolerance limits. Their phase relations are not defined as they are not synchronized in lack of a synchronizing network. Note that this applies only for higher hierarchical levels because the 64 kb/s signals and the 2 Mb/s frames are synchronized to each other. Typically, a single equipment is responsible for PCM encoding all



Midriser, *A* Midstep, *μ-karakterisztikával*

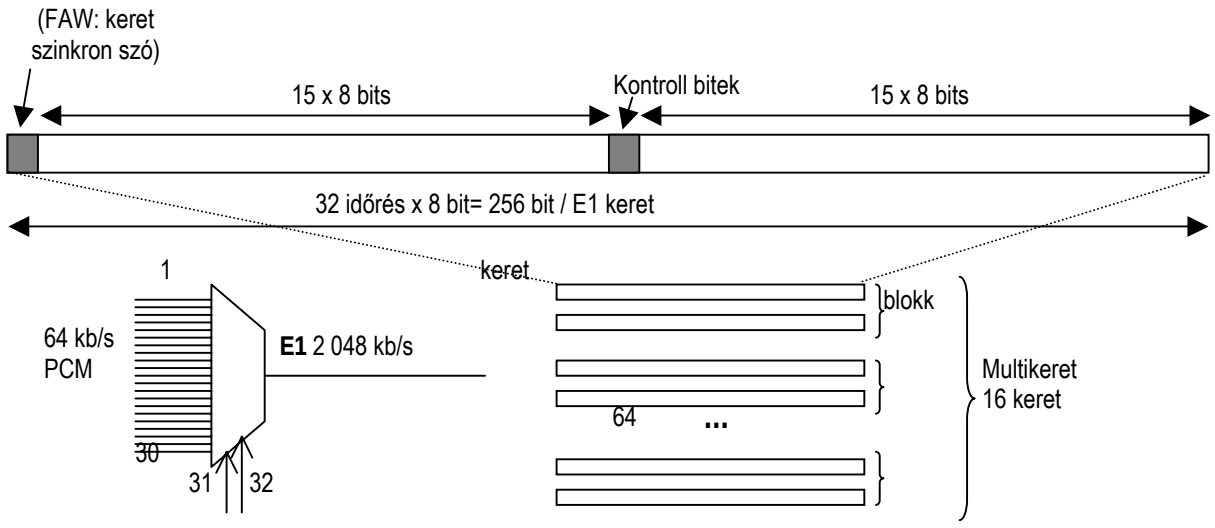


	Nominal bit rate [kb/s]	Tolerance [ppm[1]]	Line coding	Half peak voltage (V)	*a* (dB/km)	Frame size [bit]	input bit / output frame channel
E1	2 048	±50	HDB3 - High Density Bipolar Coding, limited to 3 zeroes	2,37 vagy[1] 3	6	32×8=256	8
E2	8 448	±30	HDB3	2,37	6	848	205(+1)
E3	34 368	±20	HDB3	1	12	1536	377(+1)
E4	139 264	±15	CMI - Coded Mark Inversion. Coded sign change	1	12	2928	722(+1)

represents two PCM samples per input voice channel. The Figure showt the structure os the frame, block and multiframe.

The multiframe includes 16 frames, corresponding to 8 blocks, and thus requires a transmission time of 2 ms. Within the multiframe comprising 16 of the 16th octet, the first 4 bits provide the multiframe synchronizing word, the 6th stands for error signalling and also provides 4 bits for each of the 16 samples of the remaining 30 channels. This allows for signalling transmission: half an octet signalling per multiframe takes care of 16 octets of voice, in any of the 30 channels. This results in a 2 kb/s signalling bit rate per voice channel.

Each frame starts with a frame synchronizing signal, helping the receive side to localize the frame boundaries.



Frame synchronization

- The zeroth byte of each block is a synchronizing signal having the following form:

0.	1.	2.	3.	4.	5.	6.	7.
X	**0**	**0**	**1**	**1**	**0**	**1**	**1**

- The second frame of the block begins with the inverted frame synchronizing signal. However, only the second bit marked in the Figure is monitored by the receive side that should be 1.
- Synchronization takes place by investigating the synchronizing signal at the beginning of each frame. If it is correct then synchronization is maintained as shown by the status S of the Figure below.
- If the synchronizing signal is not found in one instance then one bit may be in error, but it is also possible that synchronization has been lost. The system then passes into state A of the status diagram. If the next synchronizing signal is correct then the system returns to S, but if not then status B comes into effect. Should further on no correct synchronizing signal appear then the system goes over to the searching operation mode.
- In status K, the channel is monitored bit-by-bit, and an attempt is then made to adopt the synchronizing signal. If this is successful then status C prevails. Naturally, it may happen that bits identical with the synchronizing signal are present within the bit sequence. This means that a false frame alignment signal has been found instead of the valid signal.
- In status C, the system is waiting for the next frame, and the second bit of the frame is investigated. If this is 1, then D follows. Otherwise the system returns into K after a delay of T. (This T is not the frame time.) Synchronization takes only place if, in status D, one more correct frame synchronizing signal arrives at the correct place.
- The algorithm has protection against false frame alignment (δ=2). This means that the synchronizing signal has to be found at the correct place twice in succession in order to get back into the synchronized state.
- There is also protection against loss of frame alignment (α=3): the synchronizing signal is checked three times before loss of synchronization is ascertained. This is necessary because the synchronizing signal can be impaired by bit errors too.

E2: Secondary digital section (secondary hierarchy)

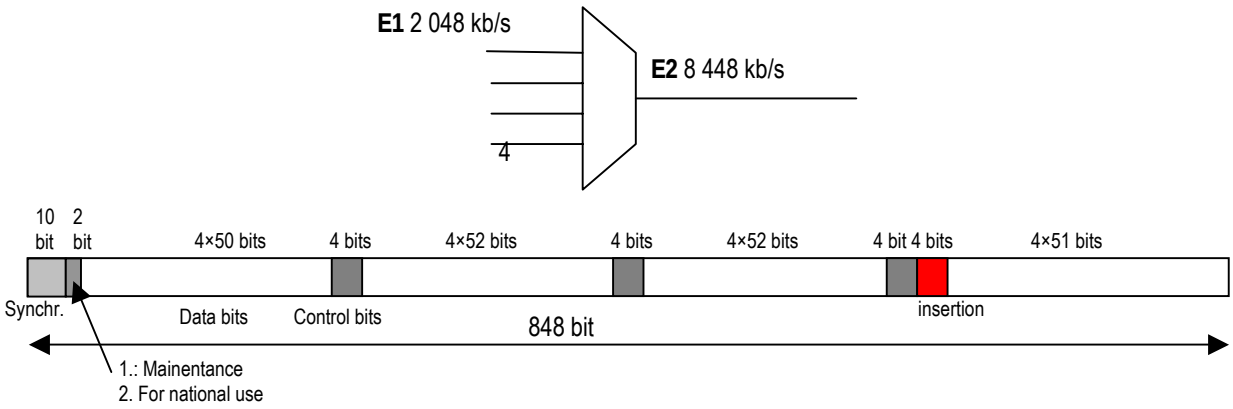
According to the Figure below, four primary E1 sections are comprised in a secondary E2 section. The E1 sections being synchronized to independent clock signals, their frequencies are different from each other and from the nominal frequency. Further, the clock frequency of section E2 may differ, within the specified

tolerance, from the nominal clock frequency. This is why it is necessary to align the frequencies of each of the four channels.

The frame structure of the secondary multiplexing hierarchy is shown in the Figure above. Following the frame alignment word and the two maintenance bits, the first bits of all channels are appearing in the bit sequence, then the second bits of all channels, and so on. The four complete E1 bit sequences, together with the synchronizing word, appear without taking into account the frame structure at the El level. This means that the E1 signals as continuous bit sequences are multiplexed together bit-by-bit to compose an E2 signal.

Bit-rate justifications

The data bits are interrupted by four-bit control bit groups .In each of these three groups, one bit is attached to each E1 input channel. This shows whether in the output E2 frame in question, one bit has or has not been used for the chosen E1 input sequence. Following the third group, majority voting in each channel will select



the channel bit sequences into which excess bits have to be inserted. For this operation, the suitable bit of the next 4-bit group is used. Bit insertion provides the possibility to equalize small differences of the channel bit rates (i.e. differences in the multiplexer clock signal frequencies). There is no possibility to insert two bits into a single channel, even if these bits are not utilized for the other channels.

The length of the complete frame is 848 bits, and from these, the number of bits transmitted in each channel is 50+52+52+51(+1)=205(+1). The required bit rate is adjusted by the frequency of inserting the excess bit. The 205-206 bits transmitted per frame and per channel assigns an interval comprising the bit rate of the individual

channels. The minimum and maximum bit rate bordering this interval are calculated as follows.

$$f_{\min}^{E2} = 205 \cdot f_k^{E2} = 205 \cdot \frac{8448 \cdot 10^3 \,\text{bit/s}}{848 \,\text{bit/s}} = 2042{,}26 \,\text{kb/s}$$

$$f_{\max}^{E2} = 206 \cdot f_k^{E2} = 206 \cdot \frac{8448 \cdot 10^3 \,\text{bit/s}}{848 \,\text{bit/s}} = 2052{,}22 \,\text{kb/s}$$

For the nominal bit rates E1 and E2, the insertion factor is given by η=0,42. This means that on the average, 0.42 bits are inserted into a frame, i.e. in the given output E2 frame, 1 bit is inserted into the given E1 input channel in 42 per cent of the frames.

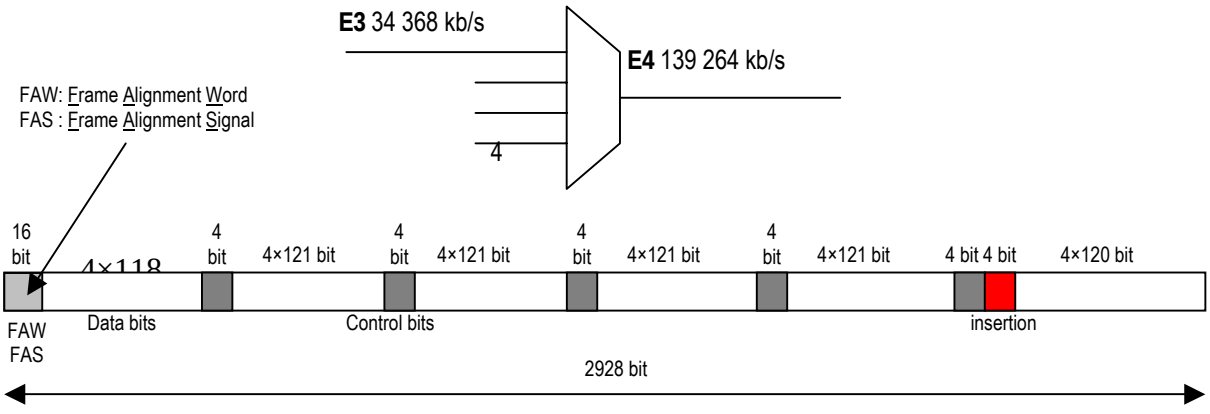
E3: Third order digital section (tertiary hierarchy)

Apart from the parameters, this level is identical with the preceding level. Here 377(+1) bits per channel are transmitted, the frequency limits being as follows:

$$f_{\min}^{E3} = 377 \,\text{bit} \cdot f_k^{E3} = 377 \,\text{bit} \cdot \frac{34378 \cdot 10^3 \,\text{bit/s}}{1536 \,\text{bit}} = 8435{,}375 \,\text{kb/s}$$

$$f_{\max}^{E3} = 378 \,\text{bit} \cdot f_k^{E3} = 378 \,\text{bit} \cdot \frac{34378 \cdot 10^3 \,\text{bit/s}}{1536 \,\text{bit}} = 8457{,}75 \,\text{kb/s}$$

E4: Fourth order digital section (quaternary hierarchy)



There are more insertion bit controls and the frame is also longer compared with the previous levels. The frame includes 2928 bits, and from these, 722(+1) bits are assigned to an input E3 channel.

Pros and cons of PDH

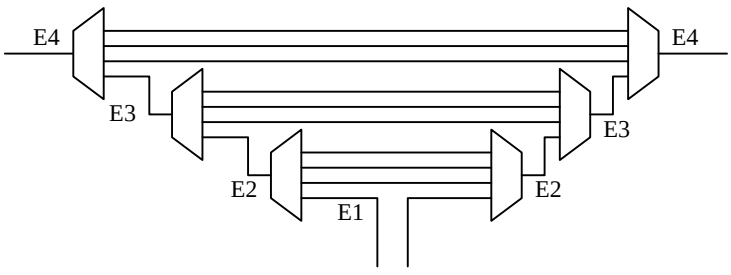
- Multiplexing is carried out bit-by-bit
- The European, Japanese and American versions are different.
- The bit rates of the individual devices may differ from the nominal bit rate, and in spite of this, the system will operate without disturbances.
- The synchronizing signal has not to be distributed in the network (there was no possibility to do this at the time the procedure was invented).
- .Frames have to be provided at each level. Let us have an example. A small village has to be supplied with an E1 signal by utilizing the nearby installed E4 line. All levels of the hierarchy have to be covered.
- There is no sufficient place for transmitting the operating/maintenance informations and possibly other kinds of informations.
- It is not easy to provide protection.
- In transmissions using modems, the useful bandwidth is limited.

2.2.1.3. SDH

ANSI conceived the SONET system (Synchronous Optical Network), primarily for transatlantic links. Subsequently, ETSI has specified SDH (Synchronous Digital Hierarchy). Finally, ITU-T (at that time, CCITT) conceived a system combining SONET and SDH, also designated SDH. The first Recommendations G.707, G.708 and G.709 have been published end of 1988. At first, it was used as the transport network of PDH networks, but it is excellent also for data transmission.

Structure of SDH networks

Principle building blocks of SDH/SONET networks are the following.



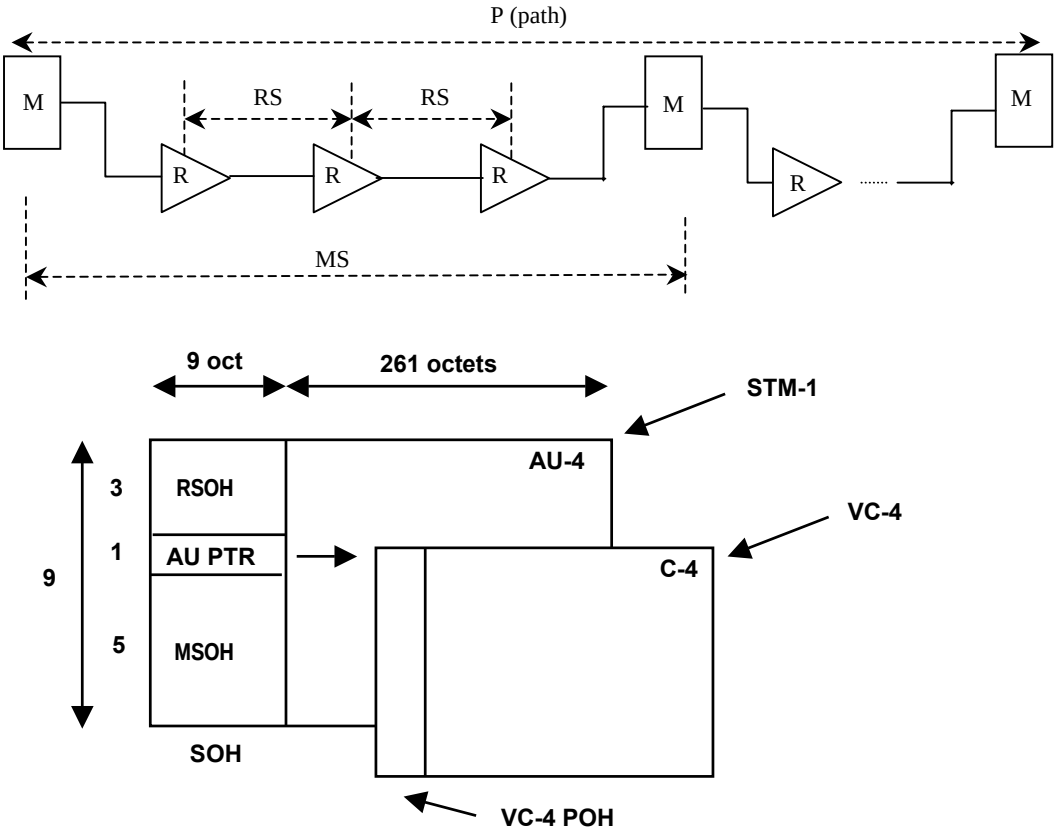
- Single mode fibre (e.g. ITU-T G.652) for connecting the netework devices. In some cases, this is substituted by a multi-mode fibre or a copper cable, for spanning short distances.
- Regenerator for signal regeneration, amplification and parity control.
- ADM (Add and Drop Multiplexer) for branching lower bit rate containers (e.g. VC-4, VC-12) from a high bit rate sequence (e.g. STM1, 4 or 16).
- DXC (Digital Cross Conect) for digital rearranging virtual containers of several multliplex sections. It is e.g. suitable for interconnecting two rings without restriction, and creating several topologies: hierarchical rings, rings connected by a mesh, a mesh netework, etc.

The Figure below shows that the section between two signal regenerators is called Regenerator Section RS. At higher levels, only sections between multiplexing equipments are distinguished (Multiplexer Section MS) while on high level, paths P are dealt with.

The frame organization corresponds to this arrrangement. Container C-4 includes 9 lines, each containing 260 octets. The useful data are placed here continuously for each line, e.g. the PDH signal at level E4, the ATM cells or the IP packets. Also, the lower level containers are also placed in C4. Additionally, the lower level carrier units are also placed in C-4. All carriers (C-x containers) are supplemented by a path overhead (POH), thus obtaining the virtual conainer (VC). The value of the POH is adjustsed only at the endings of the path as they contain information valid for the complete path. The VC is a self-contained unit that can be placed to an arbitrary point according to the higher level frame organization but can be unambiguously located by using pointers. In our Figure, the location of VC-4 within the STM-1 frame is shown by pointer 4 PTR.

STM-1 is the synchronous transport module at the first level, and comprises 9 lines x 270 octets. From these, the first 9 columns (81 octets) provide the Section Overhead SOH. The last 5 lines (45 octets) of the SOH is the Multiplexer Section Overhead MSOH that is responsible to transport the surplus information between multiplexers while the Regenerator Section Overhead RSOH is intended to transport the surplus information between regenerators. The fourth line of the STM-1 overhead is the pointer showing the location of the VC-4 payload within the STM-1.

The Figure below (ITU-T G.707) is an illustration showing how individual bit-rates can be connected to the SDH system.

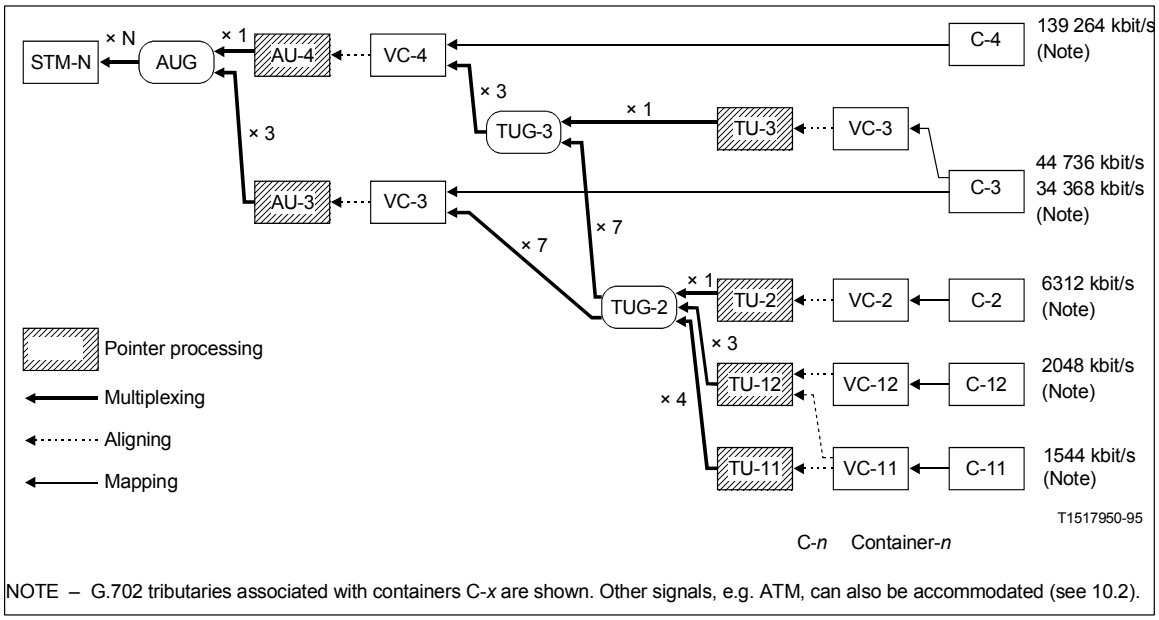


Note that this Figure shows both the ANSI SONET and the ETSI SDH “paths”. For example, AU-3, TU-11 and TU-2 are used mainly in SONET systems while TU-3, TUG-3 and TU-12 are used mainly in ETSI SDH systems.

The hierarchy of SDH and SONET is detailed in the Table below.

STM-64	9 953.28 Mb/s		OC-192 STS-192
STM-16	2 488.32 Mb/s		OC-48 STS-48
STM-4	6 220.08 Mb/s		OC-12 STS-12
STM[1]-1	155.52 Mb/s		OC-3 STS-3
	E4 139.264 Mb/s	44.736 Mb/s T3	OC[1]-1 STS[1]-1
	E3 34.368 Mb/s	6.312 Mb/s T2	
	E1 2.048 Mb/s	1.544 Mb/s T1	
	64 kb/s	DS[1]0	

By multiplexing four STM-1 signals octet-by-octet, the STM-4 hierarchy is obtained, and by multiplexing 16 STM-1 signals we obtain the STM-16 level, according to the Figure below.



NOTE – G.702 tributaries associated with containers C-*x* are shown. Other signals, e.g. ATM, can also be accommodated (see 10.2).

Note. In Rec. G.702, the components are related to the C-x containers, as illustrated. Other signals, e.g. ATM, can also be included.

In STM systems, an excellent failure protection method is available which is presented in more detail in Sec. 4.2.

Pros and cons

One of the chief advantages of the ITU-T SDH system is the fact that it may be the first compatible system used worldwide. A further advantage is the extremely high bit rate transmitted by the system (e.g. nearly 10 Gb/s with STM-64). When used in conjunction with DWDM, even much higher rates can be handled. SDH is perfectly suitable to multiplex and transport the traffic of PDH networks, and can be used also for data transport and for leased lines. It follows further from the synchronizing feature that a low level container, including also its content, can be accessed at any higher hierachical level. However, a disadvantage of SDH is the necessity to establish a synchronization network.

2.2.2. Subscriber access networks

(Editorial Chapter)

The lowest layer of the network has special significance as this is the most extended so in this layer, the use of economical solutions is of utmost importance.

In the years 1900, emphasis was placed to introduce economical solutions in this network layer**.** However, the last couple of years witnessed fundamental changes inasmuch as the mechanisation introduced in network development has drastically reduced the building costs. On the other hand, several services required the realization of broadband links reaching the subscriber homes. The most important of these services is the cable television that will reach all homes within a short period. The network structure of cable television is presented in Sec. 4.7.

Another novel requirement is the surfing over the world wide web requiring special network structures and broadband accessibility. These topics are detailed in Secs. 4.11 and 4.12. Finally, the technologies applied in diverse network layers have, in the course of times, amalgamated so the principles presented in Secs. 1.9 and 4.3 now cover all aspects that were originally planned to be included in this Chapter. A futher outcome of the development was marked by the digital subscriber access that is either based on an ISDN network or requires a digital subscriber line (xDSL).

In course of writing this book, all these items necessitated substantial changes so that the inclusion of the original chapter seemed not to be justified. It is hoped that the Sections referred to will convey the required informations.

2.2.3. *Building* Elements of Cable Television Networks

Gábor Mátay, author

Sándor Stefler, reviewer

The building up and operation of cable television (CATV) networks is described in chapter 4.8. The present modern, two-way CATV networks regarding *both their techniques and technologies are mixed systems*. This means that both digital and analogue techniques occur at program distribution and HFC (**h**ybrid **f**iber **c**oaxial: CATV with optical and coaxial cables) technology is used in their distribution network [2.2.3.1].

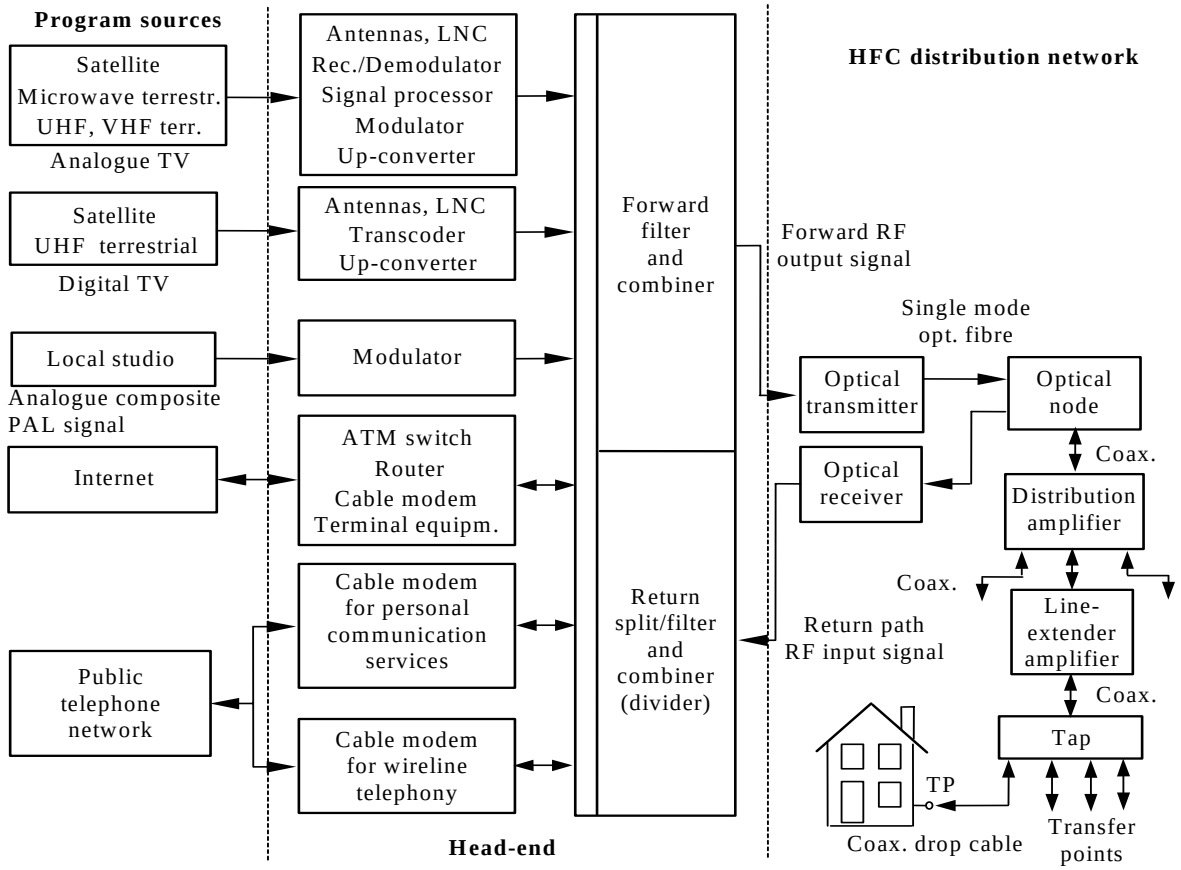


Figure 2.2.3.1 The functional block diagram of head-end and the elements of HFC distribution network.

- the CATV system has Internet and telephone services;
- in the optical part of distribution network connected to head-end space division multiplexing (SDM) is used (two, each other independent optical fibres are applied for information transmission in forward and return path);
- FDM[1] procedure is applied for multiplexing the channels used for data transmission.

Analogue and digital program channels in one satellite transponder bandwidth radiated from satellite get into of the input receiver/demodulator unit in the frequency band of low noise converter (LNC). After signal processing on the output signals of receiver/demodulator the signals is carried to modulators (remodulation). The modulations achieved in IF band are AM-VSB in case of analogue- and 64QAM for digital TV programs. Only one up conversion comes after the appropriate modulation, which transposes the channel of 8 MHz bandwidth to the frequency band used for this channel in the distribution network. After the above mentioned conversions the 8 or 10 digital TV- and 8 or 10 digital radio programs placed into the bandwidth of one

[1] FDM - **f**requency **d**ivision **m**ultiplexing

satellite transponder get to subscribers' set top boxes at their TV receivers in one UHF channel of 8 MHz bandwidth. The set top box permits of selection of the wanted program (see 4.8 chapter).

The signal obtained by summing of forward path signals gets to the optical transmitter (E/O converter). The optical output signal of transmitter is carried by monomode optical fibre to the optical node. The optical signal in return path is transported by another independent fibre from the optical node to the optical receiver (O/E converter). After distribution the return path signals of RF-band at the output of O/E converter get to a central place corresponding to the service through appropriate cable modems. The optical transmitter and receiver are often regarded as parts of head-end (in different manner as the borderline is signed by broken line in Figure 2.2.3.1).

In the coaxial part of distribution network the signals of forward and return path travel to subscriber's outlet in the same coaxial cable. The signals of two-direction are separated by diplexers. Naturally the building elements in coaxial part of distribution network shown in Figure 2.2.3.1 are also able to pass the forward and return path signals.

Optical part of distribution network (trunk) and its building elements

One optical section of the HFC distribution network (for example the forward section) consists of optical transmitter (E/O converter) and receiver (O/E converter as well as optical fibre connected them. The simplified block scheme is shown in 2.2.3.2 Figure on the basis of [2.2.3.2].

The RF input signal (all program channel to be distributed after FDM multiplexing) modulates the intensity of light radiated by semiconductor laser diode, which gets into optical fibre after focusing. Through this optical fibre the light is transported to optical receiver, which converts it back into RF signal. The *whole section has to operate with high linearity* because of transmission of analogue signals. Nevertheless the optical transmitter and receiver are nonlinear devices. The instantaneous output power represented by the light of optical transmitter is proportional to its RF input current that is the output light power is proportional to the square root of RF input power. The photodiode in optical receiver at the other end of the section gives an RF output power, which is proportional to the received optical power (it operates as square law detector); so the whole section operates linearly in principle [2.2.3.3]. In fact neither DFB[2]-laser used in optical transmitter nor photodiode operates linearly that is why a predistortion is applied in order to

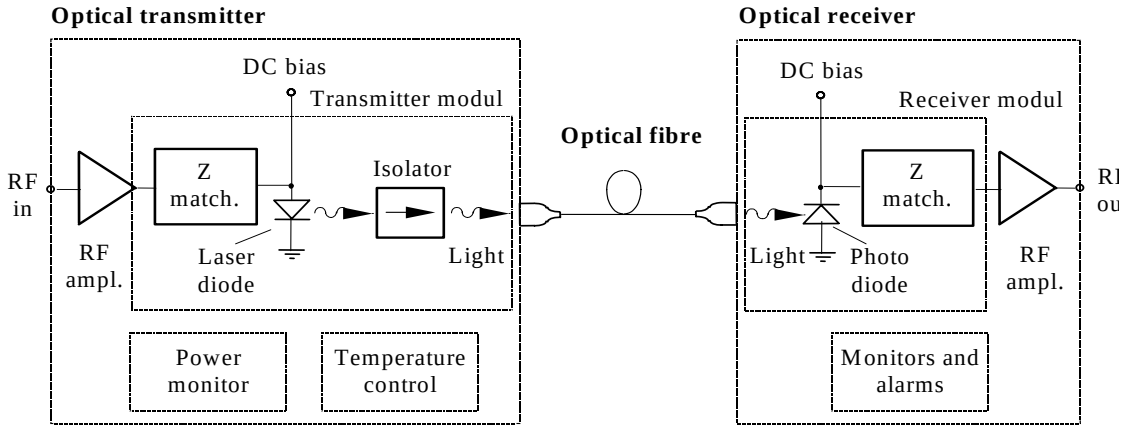


Figure 2.2.3.2 The block scheme of one optical section of distribution network.

decrease the effects of nonlinearities. By using this procedure it is attainable that the relative levels of composite second- and third-order distortion products (CSO[3], CTB[4]) be under a required value. For the better approximation to linear operation the laser diode is used in a working point of high current. The matching of diodes in optical

[2] DFB - **d**istributed **f**eed**b**ack laser: laser diode, which is used for applications of low noise and high dynamic range.

[3] CSO - **c**omposite **s**econd **o**rder: composite distortion product originating from the second-degree term of nonlinear characteristic (its exact definition is given in chapter 4.8).

[4] CTB - **c**omposite **t**riple **b**eat: composite distortion product originating from the third-degree term of nonlinear characteristic (its exact definition is given in chapter 4.8).

transmitter and receiver to RF amplifiers is realized with broadband impedance matching networks. Optical isolator at the output of transmitter inhibits the effects of reflections occurred at optical connectors on DFB-laser. Distortion products of laser diode change significantly with ambient temperature that is why an automatic temperature control circuit is applied for stabilization of its temperature. The predistortion is automatically controlled on the basis of monitoring the important parameters.

Alarm circuit in optical receiver comes into action in case of blackout of optical connection. Random fluctuations in the optical output signal of laser diodes are characterized by *relative intensity noise* (RIN[5]).

With the instantaneous output power depended on modulation of laser diode changes the wavelength of light in a small compass. This unwanted FM is called to "chirp" which causes changes of the propagation time in optical fibre in consequence of dependence on wavelength of wave velocity. This phenomenon is called to *chromatic dispersion*. The interaction among reflections at the output of optical transmitter, the input of optical receiver and occurred along optical fibre makes also worse the quality of transmitted signals. This is the case that transmission measurements are performed on the whole transmission section and in case of optical fibre actually used in given direction.

The *optical node* (see its place in CATV system in Figure 2.2.3.1), *ONU*[6] consists of optical receiver of forward path, optical transmitter of return path and diplexer. The output of optical receiver connects to the high-pass filter of diplexer, the input of optical transmitter connects to the low-pass filter of diplexer and the joining point of two filters is attached to coaxial cable of line-network as it is shown in Figure 2.2.3.3. It is important to remark, if the optical receiver has more outputs and optical transmitter has more inputs (the number of inputs and outputs is equal), then the optical node is able to feed more coaxial cables of line-network. Of course diplex filter belongs to each coaxial cable. In case of distribution network with remote power supply the optical node unit (ONU) is supplied from coaxial cable.

[5] RIN - **r**elative **i**ntensity **n**oise: it is the ratio of the mean square amplitude of the noise fluctuations per unit bandwidth (1 Hz) to the square of optical mean power. Its typical value is -160 dB/Hz in case of DFB lasers; its value made significantly worse by reflected signals from the output (see optical isolator in Figure 2.2.3.2 in order to avoid this effect).

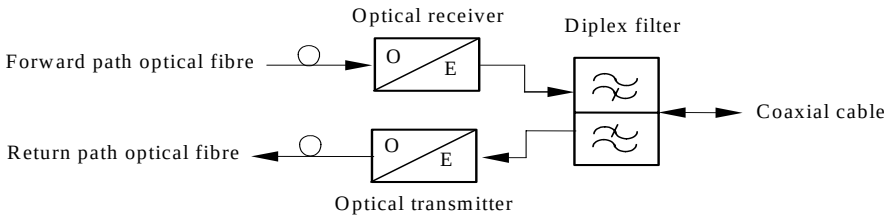


Figure 2.2.3.3 The block scheme of optical node unit (ONU).

Coaxial part of distribution network and its building elements

The coaxial part of distribution network is between optical node at the end of trunk and subscribers' outlets. Its active elements are: two-way, broadband amplifiers with and without bridger as well as amplifiers of house network.

Its passive elements are: coaxial cable, splitters (power dividers), taps, attenuators, cable equalizers, filters, power inserters and subscribers' outlets.

Requirements of ideal signal transmission for this section of transmission path are also free from linear distortion (constant amplitude-frequency and propagation time-frequency response) as well as free from noise, interference and nonlinear distortion. In the operating frequency range we have to make an effort to the best approximation of ideal signal transmission in both the forward and return path. The coaxial cable used as transmission medium in this part of distribution network is not ideal. Its attenuation rises with increase of frequency and ambient temperature. One part of passive building elements and control circuits in broadband amplifiers serves for equalization of these characteristics. Basically the operation of CATV system is limited by noise and nonlinear distortion. The *operational signal level*[7] needed to satisfy carrier-noise and carrier-distortion product ratio (C/N and C/D) specified in CATV standards must not increase and decrease along the coaxial part of distribution network. That is why the design of CATV distribution network is achieved on the basis of *unit gain conception* [2.2.3.4]. This means, that the gain of amplifiers just compensates attenuation of coaxial cable and passive elements between two

[6] ONU - **o**ptical **n**ode **u**nit

[7] Operational signal level: By definition it is the output signal level of RF amplifiers at the highest operational frequency of distribution network. (Note: The highest operational frequency is different for the forward and the return path.)

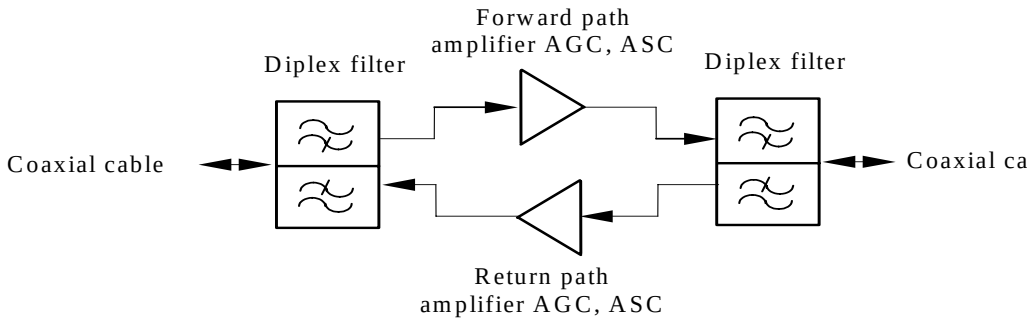


Figure 2.2.3.4 Block scheme of the two-way amplifier of line network

consecutive amplifiers. In order to decrease nonlinear distortions push-pull[8] and such kind of circuit solutions are applied in RF amplifiers, at which the signal level on amplifier element is smaller (DP[9]-, QP[10]-amplifiers) as well as the voltages and currents of working point are taken for plenty large enough compared them to the level of RF drive signal.

Simplified block schemes of the *two-way amplifier of line network* and the *distribution or bridger amplifier* are shown in Figure 2.2.3.4 and Figure 2.2.3.5

Amplifiers in both figures have changeable gain in order to adjust the exact operational level; and automatic level or gain control (ALC[11] or AGC[12]) as well as automatic slope control (ASC[13]) for continuous equalization of cable attenuation depended on ambient temperature and frequency.

.In *passive building elements* splitters (power dividers), taps and subscribers' outlets will be discussed in detail.

[8] In case of push-pull circuits (amplifiers, mixers) even-order distortion products of nonlinear characteristic disappear in principle. In practice they occur with small level in consequence of no perfect symmetry of circuits.

[9] DP - **d**ouble **p**ower amplifier: It is a circuit consisted of two push-pull amplifiers between two hybrids (functionally first one operates as power divider, second one as summator). Because of power dividing both push-pull amplifiers are driven with half power that is why carrier-third order distortion ratio improves with 6 dB compared to one push-pull amplifiers, if we assume that both hybrids are lossless.

[10] QP - **q**uadra **p**ower amplifier: It is a circuit consisted of two DP-amplifiers between two hybrids (power divider and summator). In this case one push-pull amplifier of the circuit is driven ¼ power that is why carrier-third order distortion ratio improves with 12 dB assuming lossless hybrids.

[11] ALC - **a**utomatic **l**evel **c**ontrol: It is usually realized PIN attenuator.

[12] AGC - **a**utomatic **g**ain **c**ontrol: Usually it is achieved with the change of supply voltages for amplifier element.

[13] ASC - **a**utomatic **s**lope **c**ontrol: It is automatic setting of the slope of gain-frequency characteristic for equalization of cable attenuation depended on frequency.

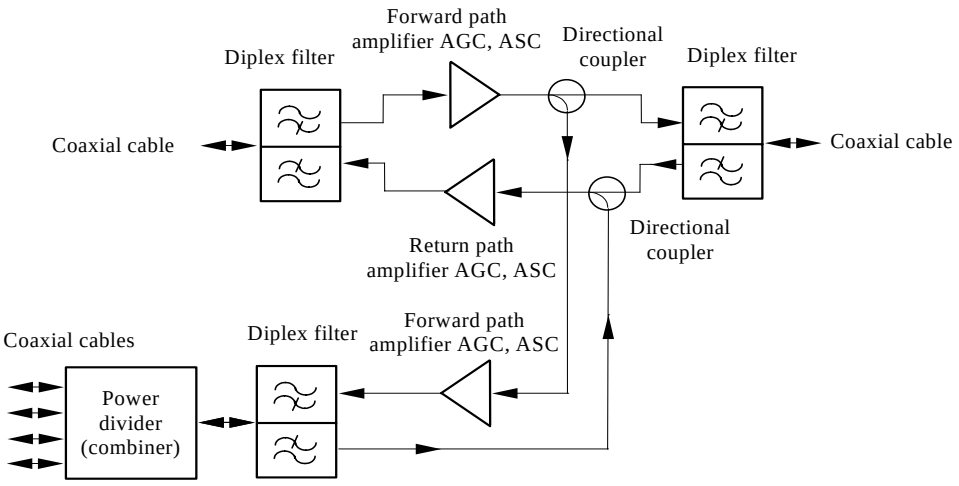


Figure 2.2.3.5 Block scheme of the two-way bridger amplifier of line network

High frequency realization of the well known from telephone technique *hybrid* is used most frequently in CATV systems as *two-way splitter* (3 dB power divider). Figure 2.2.3.6 shows its principal and practical circuit diagrams.

In case of ideal hybrid and matching loads (Z_0) the input power is divided in two equal parts for two outputs; and the attenuation between outputs (isolation) is infinite. Current on resistance of $2Z_0$ flows only in case of mismatched loads or asymmetry of coil in lengthwise. Using the outputs as inputs the hybrid operates as summator. The attenuation between input and outputs for hybrids realized in practice is typically 3,2-3,5 dB as function of frequency instead of its ideal value of 3 dB. Practical value of isolation between outputs is $\geq$ 20 dB. In practical circuit realization of hybrid shown by Figure 2.2.3.6 the tapped coil at input works as autotransformer and it is used for impedance matching; the capacitances serves for compensation of leakage inductances of coils.

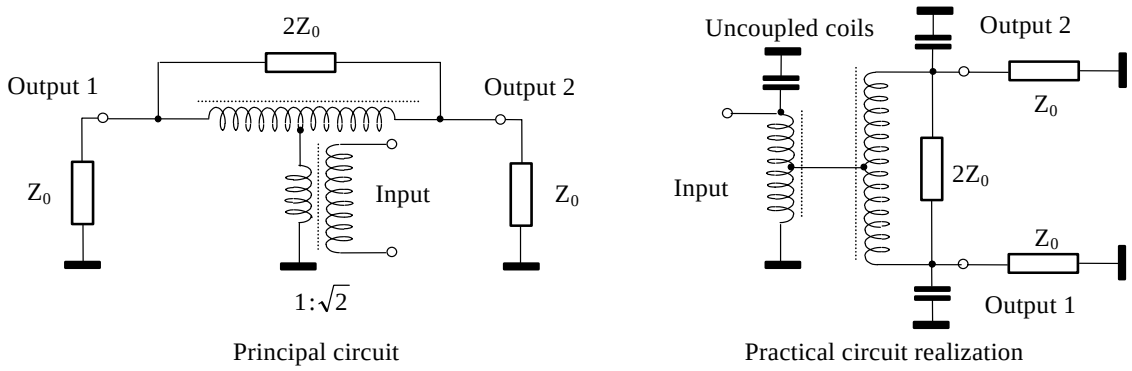


Figure 2.2.3.6 Principal and practical circuit diagram of the 3 dB hybrid power divider.

N-way splitters and taps can be realized by connecting more 3 dB hybrids or making a cascade of directional couplers. Often these two building elements are used in mixed way. Some examples are shown in Figure 2.2.3.7.

There are a lot of types of *subscribers' outlets*. The up-to-date types have one RF input, TV and FM-radio outputs and a plus two-way connector for data transmission (see its shematic drawing in the left hand side of Figure 2.2.3.8). Two things were assumed tacitly for choose of this outlet type, namely: the house network has star-structure and diplex filter is built into cable modem. This latter assumption is true for almost all of modern cable modems.

The shematic drawing and block diagram of one possible realization for subscriber's outlet can be seen in Figure 2.2.3.8.

Typical data for subscriber's outlet are following:

- Insertion loss between input and connector of cable modem is 10 dB;
- Insertion loss between input and TV output as well as input and FM-radio output is 14-16 dB;
- Isolation between connector of cable modem and TV and FM-radio output is $\geq$ 36 dB;
- Isolation between TV and FM-radio output is $\geq$ 40 dB;
- Return loss for any port of the outlet is $\geq$ 26 dB.

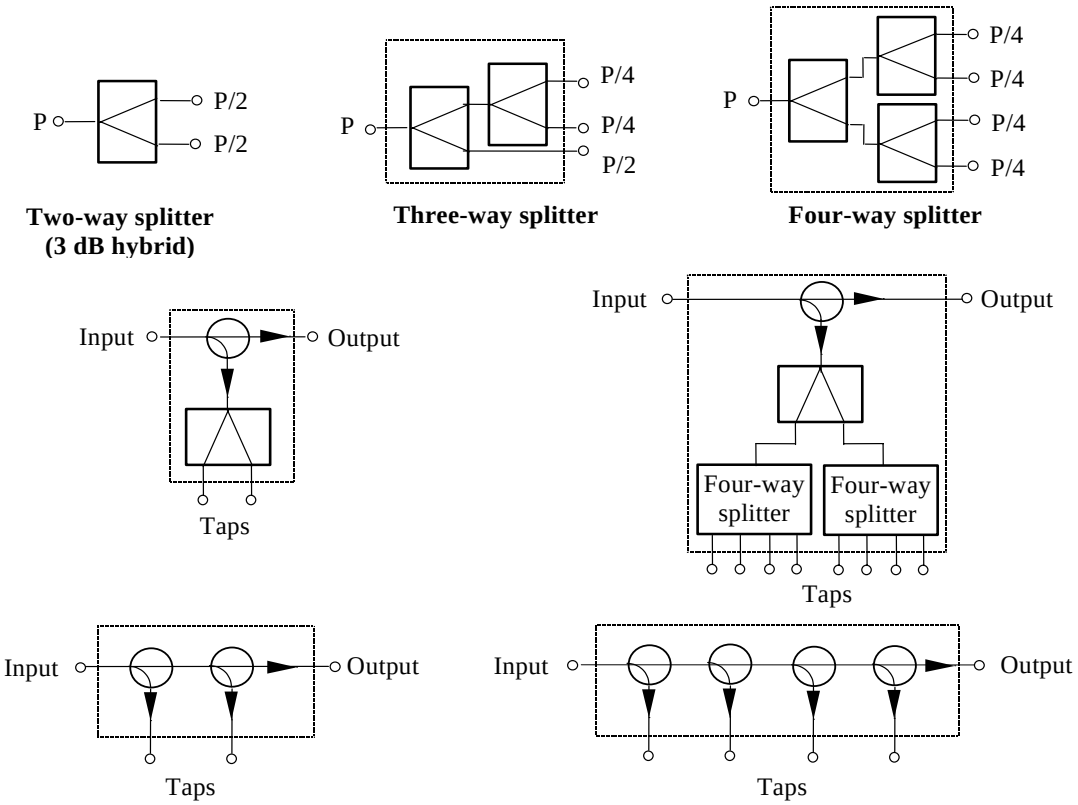


Figure 2.2.3.7. Examples for realization of splitters and taps.

The *power inserter* used for remote power supply of broadband RF amplifiers and optical node units is a simple low pass filter. Power supply unit has alternating current (AC) with frequency of 50 Hz; its voltage is about 45-60 V. There is a stabilized DC voltage source inside of all remote supplied active building elements. The DC power supply voltage of broadband RF amplifiers is usually +24 V, their current is as high as 1,2 A. At the design of remote power supply system because of high currents the voltage drops on DC resistance of coaxial cable sections have to calculate very carefully [2.2.3.5].

The *RF filters* applied in CATV systems are: lowpass, highpass, bandpass and bandstop filters. They are used for selection of program packets placed in certain frequency bands depended upon subscription rate. This makes possible only reception of certain program packets for a given subscriber very simply. They are located in closed place and inserted into the distribution network after star-point and at the input of drop cable connected to subscriber's outlet.

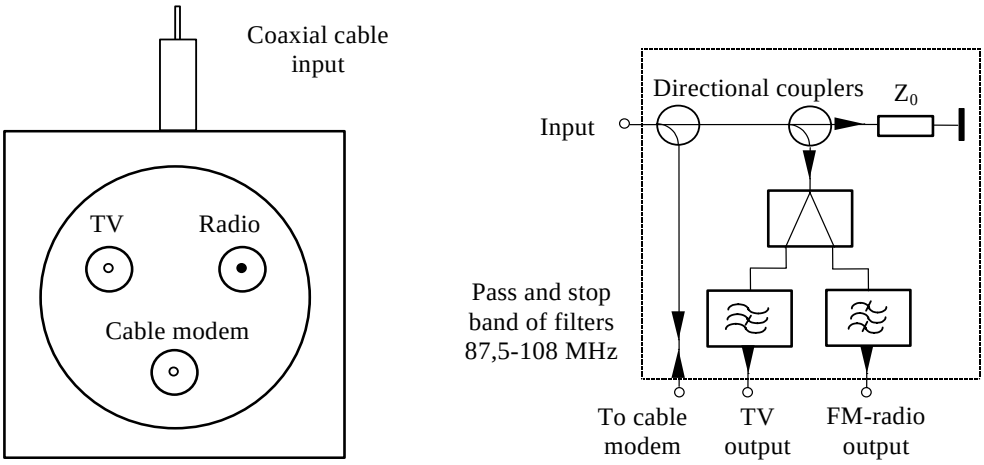


Figure 2.2.3.8 One up-to-date type of subscriber's outlet

The *set top boxes* are inserted into the system after subscriber's outlet, and they are usually put on the top of TV receiver. Their task is to allow of the reception of program channels, which cannot be received by given TV sets. For example if a TV set is unable to receive the frequency band for cable-only channels (superband, hyperband), it is possible to made it suitable for the reception of these channels by using an appropriate set top box. This set top box converts these channels to one receivable channel of TV set. In case of CATV system also distributed digital program channels and subscribers' TV sets only ready for analogue reception the set top box is a digital receiver. Its output video and audio signals in baseband are connected to the appropriate input of analogue TV set, so the digital program channels become receivable. CATV providers may also distribute program channels, which are accessible only for extra charge, that is only *conditional access* is possible to these channels. Other subscribers are unable to receive these programs because of application of special modulation or signals added to useful signals so that give the result signal unenjoyable picture and sound. The latter procedure is also called "scrambling". In CATV systems transmitted scrambled programs the set top box has such kind of circuits, which remove the added signals from result signal (descrambling) for a "command" arrived from CATV head-end so the programs of scrambled channels become receivable. Of course the command is only addressed to subscribers, who has taken in the scrambled program too. Some of set top boxes have such kind of conditional access solutions, which make impossible of watching TV programs not recommended for children. Adult subscribers themselves are able to stop these prohibitions for example by putting a key into the set top box or writing

in a password. The above-mentioned multifunctional set top boxes are usually so expensive that CATV providers give them for relevant subscribers in return for deposit.

References

[2.2.3.1] Raskin, Donald; Dean Stoneback: Broadband Return Systems for Hybrid Fiber/Coax Cable TV Networks, Prentice Hall, New Jersey, 1998, 297 pages, ISBN 0 13 636515-9

[2.2.3.2] Olson, Todd: An RF and Microwave Fiber-optic Design Guide, Microwave Journal, August 1996, pp. 54-78.

[2.2.3.3] Ciciora, Walter; James Farmer; David Large: Modern Cable Television Technology: Video, Voice, and Data Communications, 1st edition, 912 pages, Morgan Kaufmann Publishers, San Francisco, California, 1998 December, ISBN 1 55860 416 2,

[2.2.3.4] Harrel, Boby: The Cable Television Technical Handbook, Artech House, Dedham, 1985, 312 pages, ISBN 0 89006-157-2

[2.2.3.5] Solti, Miklós: Design of Cable Television Distribution Network, In Hungarian: Kábeltelevíziós elosztóhálózatok tervezése, Távközlési Kiadó, Magyar Távközlési Részvénytársaság, Budapest, 1995., 307 oldal, ISBN 963 7588 38 8

2.2.4. Wideband transmission methods

Bertalan Eged dr., author

István Frigyes dr., reviewer

In high-speed wideband distribution networks, the applied transmission methods are mainly based on the transmission of series bit sequences. By parallel connecting series transmission channels, the transmission bandwidth can further be increased. Also, the parallel connection creates redundancy that is required for high reliability links.

For wideband links operating above around 100 Mbps, practically optical transmission is exclusively used but in some cases, copper conductors are also applied for distances of up to100 to 300 meters.

In the physical layer of these high speed systems, standardized fibre channels [2.2.4.1], [2.2.4.2] are used. The layer structure and the layout of the medium coupling device are shown in the following Figures.

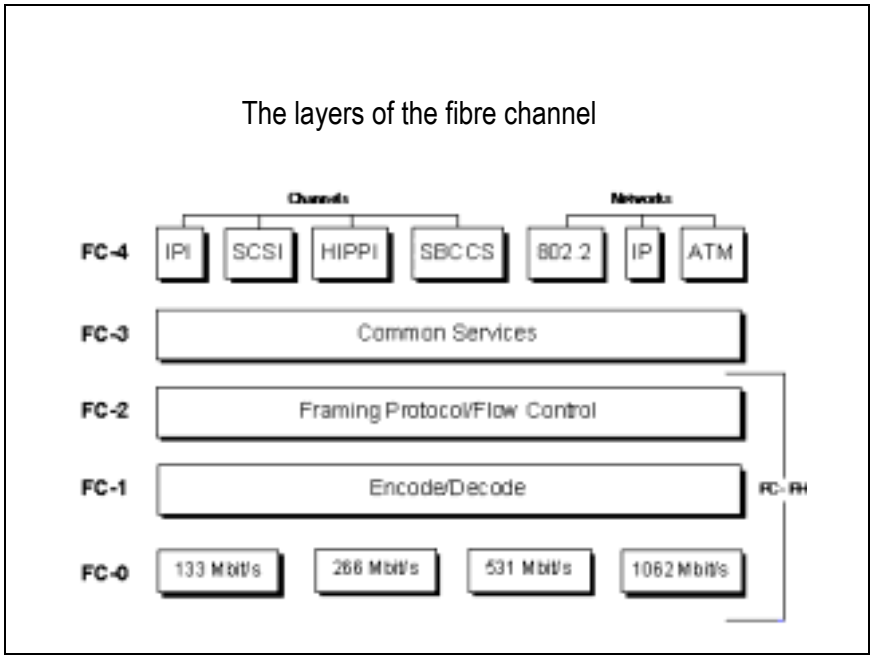


Figure 2.2.4.1. The layers of the fibre channel

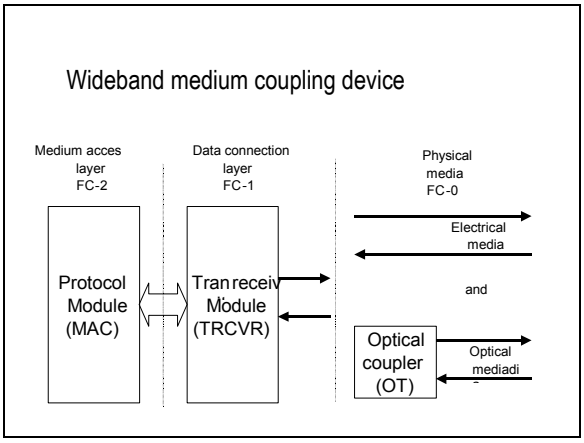


Figure 2.2.4.2 Wideband medium coupling device

Copper based transmission medium

Links are also realized by asymmetrical coaxial cables and symmetrical twisted wire pairs. Over copper wires, the series bit stream is transmitted in the baseband by pulse amplitude modulation. The complex amplitude of the discrete frequency components in this modulation as a function of the time domain pulse series is shown in the following Figures.

In wideband transmission systems, it is pracitcally more useful to consider the spectral density function envelope of the transmitted pseudo random pulses, as illustrated in the Figure 2.2.4.3.

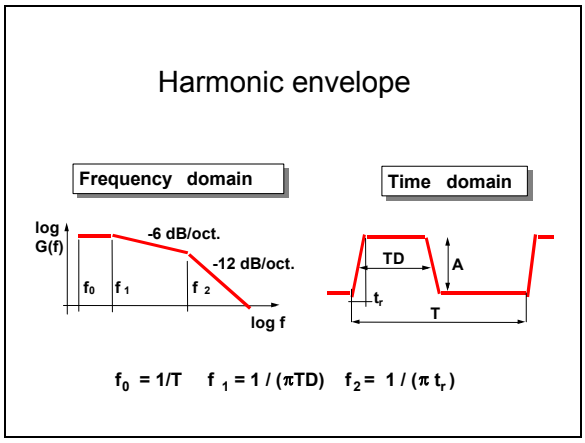


Figure 2.2.4.3. Harmonic envelope

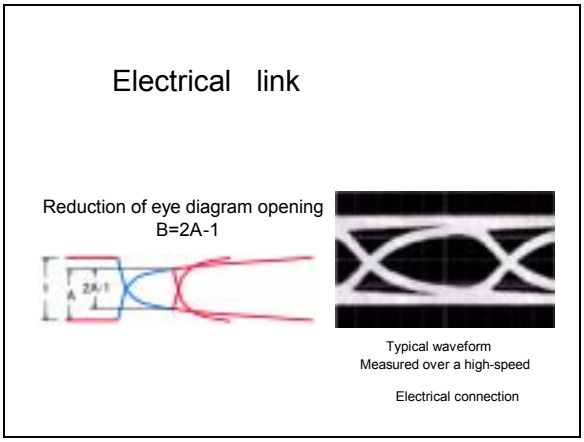
In this case, considering the properties of the transmission medium, the transmission is mainly influenced by a significant frequency dependent attenuation.

Most of the losses are due to the skin loss which is proportional to the square root of copper conductors. A smaller potion of the loss is due to dielectric losses which are generally also frequency dependent. The skin loss introduces a distortion on the time domain signals corresponding to the erfc(t) function. This can best be represented by the eye diagram of the pseudo random bit stream. The Figure below shows an eye diagram originating from a high speed signal transmitted over a coaxial cable.

The Figure 2.2.4.4 shows clearly the closure of the eye diagram which has the effect of deteroriating the bit error ratio. In addition to the attenuation, the applicable bit rate is also governed by the frequency dependence of the dielectric loss and the dielectric constant. Because of dispersion, the frequency dependence introduces inter-symbol crosstalk, that has also the effect of deteroriating the bit error ratio. Some manufacturers offer compensated cables specially developed for wideband links that are capable of establlishing longer than conventional coaxial links. However, in case of longer distances, the cost of these special compensated cables is of the same order as the cost of optical links, so for higher distances, these have not been widely used.

Optical transmission medium

In the case of wideband digital links operating in the optical domain, pulse amplitude modulation of the light ray is applied. The highest distance that can be spanned is limited by the signal-to-noise ratio required at the receiver detector input, and this is governed practically by the optical cable attenuation. By using multi-mode or mono-mode cables with different attenuations, links over distances from 300 m to 200 km can be realized.



Closure of the eye diagram
Typical waveform measured on a high-soeed electrical link

Figure 2.2.4.4. Electrical link

Optical modules

The unit responsible for pulse amplitude modulation and demodulation in wideband optical links is designated optical module or optical transceiver, in short OT. The unit has two functions. In the transmit direction, amplitude modulation of the laser source by the electrical pulses, using the on-off key procedure (OOK). In the receive direction, amplification and comparation of the photo detector signal, further generation of an output signal corresponding to the consecutive electrical interface.

In the most widely used OT realizations that are compatible with products of individual manufacturers, differential, positive emitter coupled logic (PECL) electrical interfaces are used. Electrical connections are provided by 9 terminals for connecting following signals: differential transmit and receive lines (RX+, RX-, TX+, TX-), two grounds (GND, GND), separate power supply connections for the transmit and receive branch (VCC-RX, VCC-TX), and a signal detect (SD) signal for sensing the receive side light ray.

The standard device is designated 1x9 OT, referring to the number of pins and the layout of the standard encapsulation. For connections at the optical side, SC type optical connectors are used, comprising both the transmit and the receive side connections, thus forming a dual or duplex SC connector. These kinds of optical modules are now manufactured up to the 2.5 Gbps bit rate range

The optical module comprises the transimpedance amplifier to provide matching between the electrical modulating pulse and the laser diode, further the wideband low-noise amplifier between the detector photo diode and the comparator generating the electrical receive signal. Quality parameters of the optical module are the following: optical power at the detector input required for the given bit error ratio, further the life span of the laser diode, an extremely important parameter determining the operating hours without maintenance (replacement).

The next phase of the optical module development is marked by the introduction of the so-called mini or Small Form Factor (SFF) OT's. The dimension of this OT allows to accommodate two OT's in the place of a conventional single 1x9 OT, thus further enhancing the degree of integration. The improved OT has 2x5 electrical connections, comprising also a TX disable signal, to be used for completely switching off the laser diode and the transmit circuits. At the optical side, LC type connectors are used. These are suitable for connecting with half-width connectors

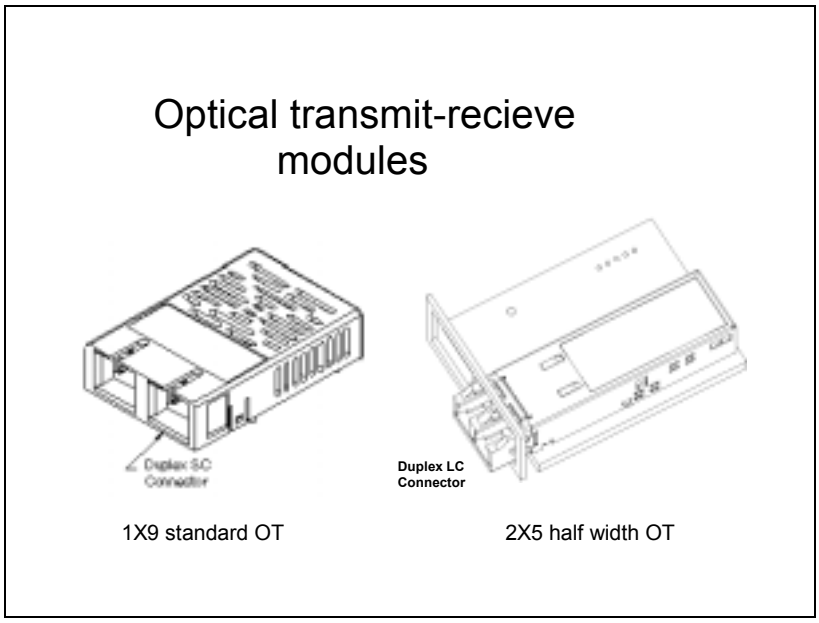


Figure 2.2.4.5. Optical transmit-receive modules félméretú – half dimension

the two optical fibres used for duplex signal transmission to the laser source and the photo detector.

Series converter

The next element of the wideband transmission system is the transceiver, providing also connection to the upper protocol layers. This provides the series interfaces needed for connections to the given transmission medium. In the case of an electrical, copper based medium, this provides direct connection, while in the case of an optical medium, the connection is established via the optical transmit-receive device OT. For distinguishing, the transmit-receive device is designated as electrical transmitter-receiver.

One of the most important tasks of the transceiver unit is the series-parallel conversion. At the series output, the clock signal frequency may be as high as several GHz so a parallel interface of lower bit rate has to be provided to face the units implementing the upper protocol layers. The format of this interface may be either 1x8 byte or 1x16 word.

Most transceivers used at present have also a channel encoding function that is based on the 8Bit/10Bit encoding, as patented by IBM. This is a static block encoding procedure in which a 10-bit code word is assigned to each 8-bit data word. By suitably choosing the codewords, the long-time average will be zero, even for arbitrary data sequences (DC compensation). The implementation of transmit or receive circuits is thus facilitated. Also, an error detection function can be applied, improved bit synchronization is possible, and code words transmitting control information required to maintain the connection are used.

The series-parallel conversion at the transmit side with the available clock signals is a relatively simple task. At the receive side, the transceiver has the task of clock recovery and regeneration. This requires at the transmit side clock signal sources having sufficiently small phase noise (in the time domain: jitter). The serial-to-parallel conversion of the received data at the parallel interface is carried out with the regenerated receive clock signal, or, when using more up-to-date devices, by synchronizing the data to the clock signal. In the latter case, syunchronization is carried out by using a first-in first-out (FIFO) memory, which has a capacity set by the frequency tolerance of the clock signal oscillators.

The next group of transceiver tasks is the implementation of link control functions. These functions involve two signals, both at the transmit side and at the receive side.

Protocol controllers

By utilizing the transceiver functions, the data entered into the transceiver will appear at the output of the receiver output at the other end. Also, by using the control signals, a point-to-point link can be established and maintained, further its status can be controlled.

The protocol drivers are connected to this interface, and establish the transport mechanism by using the signaling protocol. Also at this level, the frame organization needed for data exchange between ports, and the control of the data transmission sequence is carried out. The protocol controller should have an interface corresponding to the connected channel or network source. This means that individual protocol driver units should be used for each connection.

References

[2.2.4.1] X3T9.3 Task Group of ANSI: Fibre Channel Physical and Signaling Interface (FC-PH)
[2.2.4.2] Fibre Channel Association: Fibre Channel: Connection to the Future, 1994, ISBN 1-878707- 19-1
[2.2.4.3] Gary Kessler: Changing channels, LAN Magazine, December 1993, p.69-78

2.2.5. Fundamental properties of cellular systems. Factors influencing system parameters

Sándor Imre dr., author

László Pap dr., reviewer

In mobile telecommunication systems, the radio coverage is usually based on the so-called cellular principle, both in terrestrial and satellite systems. This requires a network of base stations over the area to be covered. Each base station covers a given region called cell. The base stations are interconnected with the switching

centres over wirebound or microwave links. Upon receiving a mobile call, the switching centre establishes connection with the base station providing the most favourable link to the called party. Sooner or later, the mobile may move to a position at which another base station provides a more favourable connection. This is when the mobile will switch over to the new base station. This switch-over procedure is called handover or handoff.

In this Chapter, the layout of the cellular systems will be investigated, and the various multiple access methods will be compared.

Evidently, the areas covered by the base stations are not exactly circular, subject to terrain conditions, coverage, weather conditions etc. However, network design and modeling calclulations based on these irregularly shaped areas are rather cumbersome. This is why in course of the first step, regular cellular-shaped (hexagonal) cells over the service area are assumed, and subsequently, local data are used to refine the obtained data.

Assuming coverage with hexagonal cells allows the calculation to be carried out much more easier in a 60° coordinate system as compared with a rectangular Descartes system. In this system, the angle between the x and y axis is 60°, and the distance D of a point (ij) from the origin is given by

$$D^2 = j^2 + ij + i^2$$

Clusters and cells

It is evident that because of mutual interference, identical frequency ranges should not be applied in neighbouring cells. However, the available frequency range for any mobile system is finite so it is not possible to use separate frequency ranges for every cell. This means that an arbitrarily large area can only be covered if the frequency used in a given cell is re-used in other cells. Therefore, the cells are grouped to form so-called clusters in which each cell has a separate frequency.

Evidently, the number K of the cluster cells should therefore be chosen so that the cells using the same frequency should be as far from each other as possible, and the clusters should establish a coverage without gaps. Due to these two conditions, there is a unique restriction for K. The cluster area can easily be calclulated as the ratio of the cluster area and the cell area:

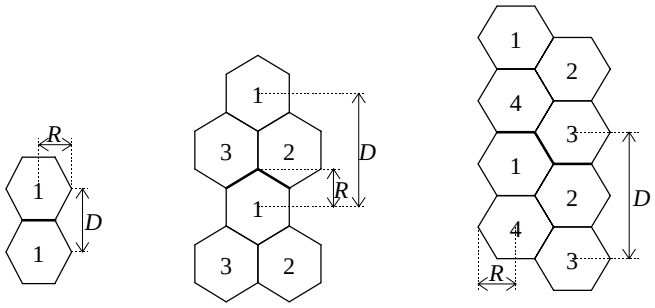


Figure 2.2.5.1. Examples for generating clusters

$$K = \frac{A}{a} = \left(\frac{D}{\sqrt{3}R} \right)^2 ,$$

where *a* is the area of one cell, *A* is the cluster area, *R* is the cell diameter, and D is the distance between clusters (the distance between cells with identical frequencies).

Considering the 60° coordinate system, let us define unity as $R\sqrt{3}$ and let us place the centre of one of the clusters into the origin. The coordinates of any cluster centre at a distance *D* from the origin are (*i*,*j*), so $K = i^2 + i\,j + j^2$.

This means that any cluster can only be generated by originating it, according to the above method, from two integers, *i* and *j,* e.g. 1, 3, 4, 7, 9, 12, 13, 16, 19, 21 etc. Figure 2.2.5.1 shows a few examples for generating clusters. Consequently, the relation between the distance between clusters and the cell diameter is given by $\frac{D}{R} = \sqrt{3K}$.

In a cellular network, there are two methods for a more efficient exploitation of the frequency band

- to apply microcells, i.e. to divide the cells resulting in so-called microcells, thus enhancing the possibility of frequency reuse,
- to apply sectors, i.e. to operate the base stations antennas with beam angles of typically 60°, 90° és 120°, instead of antennas with circular patterns. By using different frequencies in neighbouring sectors, frequency re-use can be applied even within a cell.

Interference in cellular systems

In conventional radio channels, the transmission distance and transmission quality are determined bt the transmitter power, the over-all attenuation and the

environmental noise. However, in multiple access radio systems, the main disturbance is due to interference, sometimes completely surpassing the noise effects.

Interferences in cellular systems are classified to fall into two groups:

- Adjacent channel interference
- Co-channel interference

Adjacent channel interference is due to frequiencies within a single cell, and is set primarily by the mobile phones (frequency stability, filtering, bandwidth). Co-channel interference is generated by the cells of two neighbouring clusters using the same frequency range, and depends mainly on the structure of the frequency re-use cluster, i.e. on the distance *D*. The former phenomenon is mainly equipment dependent whlile the latter is system oriented. Therefore, in the next Section, the fundamental relations related to the co-channel interference will be summarized.

Fundamental relations concerning co-channel interference

The most important suppression parameter is the co-channel supression factor

$$\tilde{a} = \frac{D}{R} = \sqrt{3K} ,$$

that is proportional to the distance *D* between clusters. This is evident because the larger the distance between cells using the same frequency the smaller will be the disturbance caused by them. On the other hand, the co-channel suppression factor is inversely proportional with the cell radius *R* because a longer cell radius results in a shorter distance between cells having the same freuency.

Let us now calculate the signal-to-noise ratio in a cell having radius *R*. This can be calculated as the ratio of the own P_j power to the total interfering power P_z falling into the band, as follows:

$$\gamma = \left(\frac{P_j}{P_z} \right) = \frac{E_s}{N_0 + \sum_{i=1}^{L} I_i} ,$$

where E_s is the symbol energy, N_0 is the single side power density of the white Gaussian noise, taken into account by a bandwidth of $\frac{1}{T_s}$, *L* is the number of interfering cells, and I_i is the symbol energy originating from the *i*th cell located at a distance D_i

Let us now consider the worst case regarding the signal-to-noise ratio: let us place the mobile receiver as far as possible from the base station, i.e. at the cell border. This also means that it is located in the closest proximity of the interfering base station. Clearly this can be assumed only for a single interfering cell as these are located, with good approximation, circularly around the reference cell. However, to simplify the calculations, it will now be assumed that this location is simultaneously valid for all interfering cells.

By applying the two-path wave propagation model, the signal power is inversly proportional to the fourth power of the distance. Assuming that all transmitters operate with the same power and the interfering cells are located circularly then the signal-to-noise ratio will be expressed by

$$\gamma \cong \frac{\left(\frac{D^4}{R^4}\right)}{L} = \frac{\tilde{a}^4}{L} .$$

This equation clearly shows that the co-channel interference can have a decisive effect on the signal-to-noise ratio.

In conventional cellular systems it is sufficient, as a first approximation, to take into account the effect of the nearest six cells generating co-channel interference. In this case, the signal-to-noise ratio and the number of cells, *K*, in the cluster, is uniquely related by the following equation:

$$K = \frac{1}{3}\sqrt{6\gamma} = \sqrt{\frac{2}{3}\gamma} .$$

K is also the segment number of the total frequency range.

Efficiency of cellular systems

Following the preceding survey of the geometry and interference relations of cellular systems, the adaptability and efficiency of various multiplexing methods will now be investigated. At present, three kinds of multiplexing methods are known:

- Frequency Division Multiplexing, FDM
- Time Division Multiplexing, TDM
- Spread Spectrum Code Division Multiplexing, CDM.

First of all, let us introduce the following notations:

B_t=1.25 MHz the total available bandwidth,

B_c bandwidth per one channel of a cell

K frequency re-use parameter (the number of cells utilizing different frequencies, the number of cells within a single cluster)

L total number of frequency ranges

Z number of TDM channels within a band

n number of channels per cell

k number of antenna segments (typically 3)

γ signal-to-interference ratio

I measuire of speech activity, expressing the percentage of time during which there is real transmitted information in th channel. Its typical value is 1/3.

Individual systems can be compared by the efficiency factors expressing the number of channels per cell. The first task is thus the calculation of n in the individual cases.

FDM analogue system

B_c = 25 kHz (FM)

K = 7 (γ = 18,6 dB)

$$n = \frac{B_t}{B_c K} = \frac{1{,}25 \cdot 10^3}{25 \cdot 7} \cong 7$$

TDM digital system

B_c = 25 kHz (FM)

Z = 3 (~8 kbit/sec channel)

K = 4 (γ = 13,8 dB)

$$n = \frac{B_t}{B_c K} Z = \frac{1{,}25 \cdot 10^3}{25 \cdot 4} \cdot 3 \cong 37$$

CDM digital system

$B_t = B_c = 1.25$ MHz (FM)

$PG = 156$

$K = 1$; $k = 1$ vagy 3; $l = 1/3$

$$\gamma © = \frac{E_s}{\sum I_i} N = 7\text{dB}\,;\ \gamma = \frac{E_s}{\sum I_i} = -15\text{dB}$$

In the CDM system, the system noise is dominant, so it is feasible to calculate the highest number of channel as follows.

Mobile station→ base station direction (in case of power control)

In the case of M mobiles, the system noise is determined by the interference generated by (M-1) mobile units, so the signal-to-interference ratio is given, with good approximation, by

$$\gamma_{m\to b} = \left(\frac{P_j}{P_z}\right)_b \cong \frac{1}{M-1}.$$

Base station → mobile station direction

$$\gamma_{b\to m} = \left(\frac{P_j}{P_z}\right)_{b\to bm} \cong \frac{1}{1{,}104M - 0{,}33}.$$

In our case, if $\gamma = -15$ dB, $M = 28$, then

$$n = Mkl = \begin{cases} 84 & \text{if } k = 1 \\ 252 & \text{if } k = 3 \end{cases}$$

Efficiency parameters

In the followins, the efficiency parameters to be calculated from the channel number will be summarized.

Area efficiency

The number of channels per unit area and unit frequency is given by

$$\eta_t = \frac{n}{R^2 \pi B_t}\left[\frac{\text{channel number}}{\text{km}^2\,\text{MHz}}\right]$$

For instance, in the case of FDM,

$$\eta_t = \frac{7}{R^2 \pi 1{,}25} = \frac{1{,}78}{R^2}\left[\frac{\text{channel number}}{\text{km}^2\,\text{MHz}}\right]$$

Service efficiency

This is expressed by the number of users who receive the service, and is calculated from the blocking probability.

$$N_f = \frac{T}{\pi R^2} M\left(n, P_B, \bar{t}\right),$$

where T is the total covered area. *R* Is the radius of the cells, *n* is the number of available traffic channels per cell, P_B is the blocking probability, and $\bar{t}$ is the average call time (in minutes), and

$$M\left(n, P_B, \bar{t}\right) = \frac{F\left(n, P_B\right)}{\bar{t}} 60$$

is the number of users serviced in one cell. Here $F\left(n, P_B\right)$ is the Erlang B formula, expressing the offered traffic $\frac{\lambda}{\mu}$ as a function of the number of channels and the blocking probability, as expressed by

$$P_B = \frac{\left(\frac{\lambda}{\mu}\right)^n \frac{1}{n!}}{\sum_{i=0}^{n}\left(\frac{\lambda}{\mu}\right)^i \frac{1}{i!}}.$$

Consider now a Hungarian example related to the precedings.

$$\left.\begin{array}{l} T = 100.000\,\text{km}^2 \\ R = 0.1, 1, 10\,\text{km} \\ \text{n} = 7 \\ P_B = 0.02 \\ \bar{t} = 1.76\text{minutes} \\ F(7;0.02) = 2{,}94 \equiv \frac{l}{m} \\ M = \frac{2{,}94}{1{,}76} 60 = 100 \end{array}\right\} N_f = \frac{100000}{R^2 \pi} 100 = \frac{10^7}{R^2 \pi} = \begin{cases} 3.18 \cdot 10^8; & R = 0.1\,\text{km} \\ 3.18 \cdot 10^6; & R = 1\,\text{km} \\ 3.18 \cdot 10^4; & R = 10\,\text{km} \end{cases}$$

2.2.6. Air interfaces and specific elements of terrestrial mobile systems

Andrea Bausz dr., author

Csaba Kántor dr., reviewer

2.2.6.1. Public cellular radiotelephone systems

First generation (1G) radiotelephone systems of the eighties are called analogue networks because of the use of analogue radio technology. These analogue systems applied solely frequency division multiple access (FDMA) as a multiple access technique, and the channel spacing usually was 25 kHz.

Second generation (2G) systems emerging in the nineties use already digital radio technology. In the following the main air interface characteristics of GSM, the Pan-European digital mobile system are briefly presented.

2.2.6.1.1 GSM air interface

2.2.6.1.1.1. Main radio characteristics

Spectrum

GSM900 is operated in the 890-915 / 935-960 MHz primary or in the 880-915 / 925-960 MHz extended band, and GSM1800 works in the 1710-1785 / 1805-1880 MHz frequency band, while GSM1900 in America is operated in the so called PCS 1900 band.

Channel spacing

The RF channel spacing of the GSM standard is 200 kHz.

Multiple access

GSM applies a combined frequency and time division multiple access scheme (FDMA/TDMA) providing 8 full rate (FR) or 16 half rate (HR) timeslots, i.e. connections per carrier.

In GSM basic services one timeslot is allocated to the user on the duplex pair of carriers. There is a shift of 3 timeslots between the transmission and reception of one user, and during that period the radio unit can be re-tuned, so it is sufficient to use only a single synthesiser radio unit in the equipment

Classes and power of mobile stations

GSM mobile stations are classified according to their maximum transmission power. There are four MS classes in GSM900: 8 W for vehicle-mounted or portable stations, and 5 W or 2 W or 0,8 W for handheld MS. In GSM1800 there are two classes for handheld mobiles, 1 W and 0,25 W. Mean power is of course the one eighth of the maximum value, because each mobile equipment is transmitting only in one eighth of the time duration of the TDMA-frame (in case of full rate channels).

Power control

Transmitted power of both the mobile and the base stations can be adjusted in fifteen steps of 2 dB, in every 60 ms under the control of the BSS. Power control keeps the transmitted power on the possible minimum level in accordance with the surrounding conditions thus the interference level can be decreased in the network.

Modulation

Modulation on the GSM radio channel is GMSK (Gaussian Minimum Shift Keying), with a modulation index of BT=0,3. It provides the transmission rate of 270,833 kbit/s per carrier.

Maximum cell radius and maximum speed

The GSM system is able to correct the Doppler shift up to the vehicle speed of 250 km/h, and able to handle the propagation delay up to the cell radius of 35 km.

Frequency hopping

Slow frequency hopping is an optional feature in GSM networks. It means that each TDMA frame is transmitted on a different carrier chosen by a pre-defined pseudo-random hopping scheme providing 217 hops per second. Frequency hopping optimises the quality of the connections, since the frequency dependent transmission errors are meant up.

Discontinuous Transmission

Discontinuous Transmission (DTX) comes into action, when the user is just not speaking during a call. The pause is detected by the Voice Activity Detection (VAD) circuit, which switches off the transmission for that time period, when there is no new information. In order not to confuse the partner with the absolute silence, his/her mobile station creates an artificial background noise based on the characteristics of the real one. Discontinuous Transmission spares the battery of the mobile station and decreases the unnecessary signal radiation.

2.2.6.1.1.2 Digital signal processing

GSM uses the LPC-LTP-RPE (Linear Predictive Coding – Long Term Prediction - Regular Pulse Excitation) method based on modelling voice creation for speech coding. It allocates 260 bits to a 20 ms piece of the analogue speech, resulting in the coding rate of 13 kbit/s.

Channel coding is the next step, first with an outer block coding, then a regrouping followed by an internal convolution coding. So after these error correction coding processes the voice sample of 20 ms will be equivalent of 456 bits (8 blocks of 57 bits each), which corresponds to the 22,8 kbit/s transmission rate.

Then interleaving is coming with separating the blocks belonging to one speech sample. Consequently in case of transmission errors it will be not one single sample with irreversible damage but rather several samples with minor injuries, since the error correction process is able to correct these ones.

Finally encryption is done with a secret key known only by the mobile and the base stations, followed by the modulation.

2.2.6.1.2 GPRS (General Packet Radio Service)

GPRS is a new packet switched bearer service based on the existing circuit switched GSM system [2.6.3]. The GPRS air interface is built on the GSM standard, however the circuit switched GSM remains untouched with GPRS introduction. For that very reason GPRS also uses the GMSK modulation method of the GSM air interface. Transmission of data packages on the GPRS air interface is done in blocks, each containing 456 bits. Radio resources are allocated to the blocks and not to the traffic channels, as in GSM. This process allows a more efficient utilisation, as GPRS allocates resources dynamically only in that case, when really there is data

traffic. Thus more users share the same physical channel and GSM and GPRS users have common access to the same radio resources. GPRS allows multi-slot operation, when a mobile station may transmit in several timeslots of a TDMA frame, furthermore GPRS allows the separate handling of uplink and downlink resources, thus supporting asymmetric traffic.

GPRS introduces three new channel coding schemes, which allow higher transmission rates in one timeslot compared to GSM (see Table 2-6-1), however provide less error correction capability. The maximum bit rate available in GPRS is 171,2 kbit/s in principle, by bundling eight time slots and using the channel coding scheme providing the lowest data protection. However in practice CS-2 coding scheme and bundling of 3 time slots DL are available from both terminal and network side at the moment, so the radio channel provides 40,2 kbit/s, which is decreased to a 30-33 kbit/s rate on the application layer level, depending on the application.

Coding Scheme	Error correction (convolutional) coding	Data rate per timeslot [kbit/s]
CS-1	yes	9,05
CS-2	Yes (puncturing, 2/3)	13,4
CS-3	Yes (puncturing, 3/4)	15,6
CS-4	no	21,4

Table 2.6.1: GPRS coding schemes [2.6.2]

2.2.6.1.3 EDGE (Enhanced Data rates for GSM Evolution)

EDGE is built on a different modulation scheme as the traditional GSM. EDGE introduces the 8-PSK modulation method, which allows higher transmission rates. In the frame of EDGE the enhanced packet switched EGPRS will support data rates up to 384 kbit/s by the speed limit of 100 km/h.

2.2.6.1.3 Terrestrial air interfaces of the third generation mobile systems

IMT-2000 (International Mobile Telecommunications), the third generation mobile standard prepared under the co-ordination of ITU, was planned as a global standard with a single air interface according to the original targets defined in the eighties. ITU specified the guidelines and the requirements only, and invited applications for the elaboration of the Radio Transmission Technologies (RTT).

After lengthy discussions and harmonisation process finally five terrestrial RTTs have been approved from the sixteen applicants, since the main regional

interest groups were unwilling to make further compromises. ITU-R, fulfilling a co-ordinator role only in the RTT standardisation process, was compelled to accept that IMT-2000 would be a family of standards instead of a world standard.

Table 2-6-2 lists the terrestrial air interfaces of the IMT-2000 family of standards. However in the ITU Recommendations [2.6.4] the single standards appear with new, uniform IMT names and abbreviations, not with their original, regionally well known names.

Though the single 3G air interface has not been realized, with this step the IMT-2000 family of standards satisfies the demand that the regionally different 2G mobile systems can be gradually developed towards 3G service capabilities by choosing the relevant air interface.

Abbrev.	IMT-2000 terrestrial RTT	3G standard	2G background standard	Standardisation body
IMT-DS	Direct Spread, W-CDMA	UMTS FDD (W-CDMA)	GSM	3GPP (ETSI, ARIB)
IMT-MC	Multi-Carrier, W-CDMA	cdma2000	IS-95 CDMA (cdmaOne)	3GPP2
IMT-TC	Time Code, W-CDMA TDD	UMTS TDD	GSM	3GPP (ETSI)
IMT-SC	Single-Carrier, TDMA	UWC-136	IS-136 TDMA	UWCC
IMT-FT	Frequency-Time, TDMA Multi-Carrier	DECT	DECT	ETSI Project DECT

Table 2.6.2: Terrestrial air interfaces of the IMT-2000 family of standards

References

[2.6.1] Mouly-Pautet: The GSM System for Mobile Communications; Cell&Sys, 1992.

[2.6.2] GSM 03.64 version 5.2.0 1998-01: Digital cellular telecommunications system (Phase 2+); Overall description of the GPRS radio interface; Stage 2

[2.6.3] Bettstetter-Vögel-Eberspacher: GSM Ph2+ GPRS: Architecture, Protocols and Air Interface, IEEE Communications Surveys, Third Quarter 1999, vol. 2 No 3.

[2.6.4] Rec. ITU-R M.1457: Detailed Specifications of the Radio Interfaces of International Mobile Telecommunications-2000 (IMT-2000)

2.2.7. Wireless office systems

Sándor Imre dr., author

László Pap dr., reviewer

Wireles local area networks (WLAN)

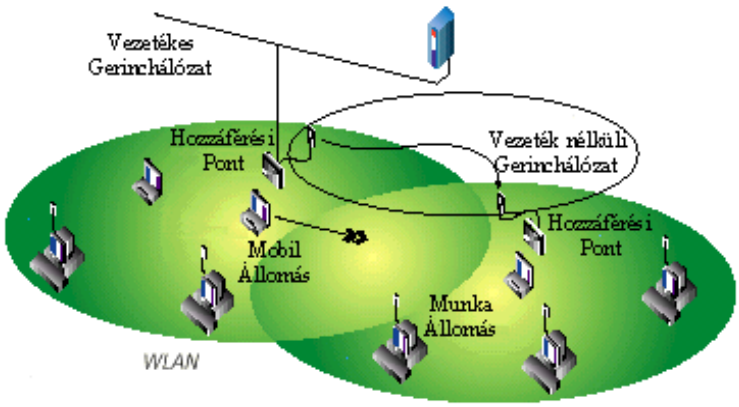
The development of computers and the increasing demand for information paved the way for data accessing methods, independently from the location.

Wireless local computer network systems (WLAN, Wireless Local Area Network) may provide an alternative solution for implementing local computer networks, or may serve to extend a structured cable system. These two solutions may jointly contribute to span the last few meters of 3G and subsequent mobile networks.

Advantages of using WLAN networks

- The cost of implementing the network infrastructure can be decreased by jointly applying wirebound and wireless network infrastsructure.
- The wireless backbone network facilitates the development of independent WLAN systems, even in those areas in which the installation of cable networks is difficult.
- WLAN networks have higher implementation costs but their life cycle costs, in the dynamically changing environment, are much lower.
- No wire link is necessary beween the access points and the workstation, so the operating cable system can rapidly be extended.
- Implementation is possible even in areas where the establishment of the cabling system is difficult.
- Re-configuration of the operating WLAN network is rapidly implemented and requires no cost, which is very important in a dynamically changing environment.
- A computer can be used more efficiently when the network can be accessed independently from the location
- Working efficiency while using a computer is enhanced by the mobility and the possibility to access the netrwork independently from the location. Data handling, loading and reading off can be directly carried out at the point of origination. This feature may be of importance in hospitals, or when managing warehouses while having information exchange with a central data base, or when information has to be exchanged within a small group, e.g. during a meeting.

There are two kinds of mobilities in WLAN systems. Total mobility is present when the terminal in move within the total area covered by WLAN is able to transmit and receive information packets. In the case of partial mobility or portability, the



2.2.7.1. ábra Building blocks of a wireless local network

terminal can be operated only in a fixed position, but at any place within the covered area.

WLAN systems can be implemented by following the existing network topologies. WLAN manufacturers may offer Ethernet, Token-ring and other well-known interfaces for their products that can be used for connecting the WLAN to the wirebound networks.

At a fixed location, the acces point or points are connected to the wirebound network by using the standard cabling. The access point, operating as a bridge or router, has the function to store and forward the data between the wirebound network and the terminals in another cell or in the own cell. End users may access the WLAN networks over a wireless LAN adapter acting as an interface between the client operational system and the transmission medium.

Considering the OSI model,the wireless local networks can be defined by stating the physical and data connection layer. In the following, main parameters of these two layers will be summarized.

The physical layer

In WLANs, basically either a radio frequency or an infrared transmission medium is applied. WLAN networks operating in the optical domain can be classified, from the operating viewpoint, to fall into two groups, diffuse systems and systems based on the direct transmission of light rays. In the case of diffuse transmission, there is no direct signal path between the transmitter and the receiver, and the transmission link is based on reflection.

For small office networks within a few rooms, the optical solution is less expensive, but for covering a complete building, a radio frequency solution is rather

used because with optical transmission, the cell could cover a single room only as the light ray cannot pass through walls. On the other hand, optical transmission is feasible in environments exhibiting high electromagnetic noises.

Spread spectrum radio LAN networks operate in industrial, scientific and medical (ISM) frequency bands that are licence free bands. This means that several separate WLAN networks may be operated at not two high distances from each other without interference. Attention should be payed to avoid interference as radio frequency signals penetrate the walls, and lthe cell area maycover a whole building. However, interference may originate from other sources too, e.g. microwave owens are operated at 2460 MHz, falling into the 2.4 GHz frequency range. In up-to-date WLAN systems, the least interfered frequency band is selected by an inherent automatic system.

LANs operating within indoor environments have to face further problems caused by multipath propagation, Rayleigh fading and absorption. Most building materials will not pass IR radiation, resulting in a coverage area of a single room only. On the other hand, RF is practically immune to absorption and reflection. Muiltipath propagation has an influlience in both kinds of system types. Rayleigh fading, a problem of RF systems, is present if the path difference between signals arriving over different paths is an integer multiple of the half wavelength. In IR systems, the effect of Rayleigh fading can be neglected becauise of the extremely short wavelength. On the other hand, the wavelength is comparable with the dimensions of a laptop so the fading may significantly change by moving the computer.

In RF LANs, spread spectrum (SS) transmission is frequently applied in two versions: frequency hopping spread spectrum (FHSS) and direct sequential spread spectrum (DSSS) transmission.

To avoid multipath transmission effects, Orthogonal Frequancy Division Multiplexing (OFDM) is applied in most up-to-date WLAN systems (IEEE 802.11b, HIPERLAN2). These proved to be extremely robust in indoor environments.

In Table 2.2.7.1, the main characteristics of WLAN systems are summarized.

Characteristic	802.11	802.11b	802.11a	HiperLAN/2
Spectrum	2.4 GHz	2.4 GHz	5 GHz	5 GHz
~Max physical rate	2 Mb/s	11 Mbit/s	54 Mb/s	54 Mbit/s
~Max data rate, layer 3	1.2 Mb/s	5 Mb/s	32 Mb/s	32 Mb/s
Medium access control/Media sharing	Carrier sense – CSMA/CA			Central resource control/ TDMA/TDD
Connectivity	Conn.-less	Conn.-less	Conn.-less	Conn.-oriented
Multicast	Yes	Yes	Yes	Yes[3]
QoS support	(PCF)[2]	(PCF)[2]	(PCF)[2]	ATM/802.1p/RSVP/ DiffServ (full control)
Frequency selection	Frequency-hopping or DSSS	DSSS	Single carrier	Single carrier with Dynamic Frequency Selection
Authentication	No	No	No	NAI/IEEE address/X.509
Encryption	40-bit RC4	40-bit RC4	40-bit RC4	DES, 3DES
Handover support	(No)[3]	(No)[3]	(No)[3]	(No)[4]
Fixed network support	Ethernet	Ethernet	Ethernet	Ethernet, IP, ATM, UMTS, FireWire, PPP[5]
Management	802.11 MIB	802.11 MIB	802.11 MIB	HiperLAN/2 MIB
Radio link quality control	No	No	No	Link adaptation

Table 2.2.7.1.Characteristics of modern WLAN systems

The data connection layer

Basic functions of this layer are the regulation of the medium access and the hiding of the physical layer from before the upper layers (transparency). The WLAN MAC protocols present new challenges as compared with conventional LAN protocols (Ethernet, Token Ring). This is explained by the fact that multipath propagation is present in an indoor radio environment, so the field strength may drastsically change within small distances, and even fall below the required level. Another problem is presented by the carrier sensing which needs a much longer time (appr. 30 to 50 µsec), possibly using up a significant part of the packet time. The CSMA protocols can therefore be used only with modifications.

A further significant difference is the necessity of applying cells for covering the area. These cells utilize differing frequencies in order to decrease the interference. Each cell can be regarded as a LAN. The access point is generally termed Mobile Support Station (MSS), and the terminal designation is MH (Mobile Host). The MSS has the function of providing connection to the wirebound networks and seizure within the cell.

In a WLAN environment, the individual communication directions have to be distinguished. This is necessary because due to the utilized applications, e.g. file downloading, the downlink communication mounts up to appr. 70 to 80 per cent of the total traffic. It is expedient to control the downlink channel completely by the MSS in order to provide efficient bandwidth utilization and quiality of service (QoS). This will provide seizure of the downlink channels according to the subscriber requirements. The uplink is much more complicated because it is not possible to have constant channel reservation because of the dynamic change of the number of users. This is emphasized when preparing the MAC protocol. For instance in the case of EIII802.11, 20 basic requirements have been defined. The most important requirements are the maximization of the throughput with minimum delay, and securing a fair access for all subscribers.

Concerning the network implementation, it should be noted that there are significant differences between ad-hoc and infrastructure networks. In the latter, the MSS is able to organize the channel seizure and the schedule in a centralized manner, by meeting the subscriber requirements, efficiently utilizing the bandwidth. In the ad-hoc case, there is no centrall lschedule so the access protocol should be based on coincidence.

All multiple access protocols applied in infrastructure WLANs have the following common properties.

1. For the MHs willing to transmit the data packet, the contest is announced (directly or undirectly) by the MSS.
2. The stations willing to send the packet are competing for the channel, e.g. with CSMA.
3. The channel is seized by the MS for the successful station.
4. The stations transmit their packets in a coincidence-free period.
5. The MSS transmits a direct acknowledgement ACK for the received packets

Wireless Local Loop (WLL)

According to definition, wireless local loops are radio systems providing telephone connection in a domestic environment without wirebound access. Figure 2.2.7.2 illluistrates the appllication environment of WLLs. The homes are interconnected via a distribution point by a local loop. Typically, the local loop is

implemented by a copper balanced pair, the switch and the distribution point being interconnected by a trunk cable.

The WLL is used for replacing the local loop section by a radio section, affecting onltthe part between the distribution point and the house, the rest of the network remaining unchanged. The use of the radio section has several advantages as it can be implemented rapidly at lower costs.

In the following, two WLL types will be presented.

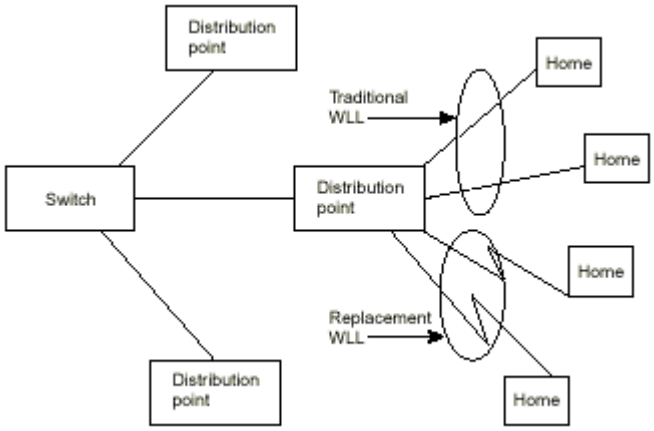


Figure 2.2.7.2. WLL and its environment

Telephone based systems

A wide range of telephone based WLL systems provides transmission bit rates from 9.6 kb/s to appr. 384 kb/s. These systems are usually classified to fall into five groups: digital cellular systems, analogue cellular systems, personal communication service (PCS), cordless telephones (e.g. DECT), and private systems.

WLL advantages:

- Wide environmental operating range
- Provide both voice and data transmission

WLL disadvantages:

- Transmission bit rates do not allow the implementation of high speed connections

Video-based systems

In the latest experimental WLL systems, TV signals and wideband data are also transmitted. These systems are designated as microwave video distribution systems (MVDS) operating in the 4 GHz band over a frequency range of 2 GHz. These have typically asymmetrical terminals, with input bit rates of 500 Mb/s and output bit rates of 20 kb/s.

Advantages:

- Required input transmission bit rate
- A wide range of video, telephone and data services withein a single access network
- Low cost as compared with wirebound solutions

Disadvantages:

- Due to the low output bit rate, duplex video applications are cumbersome (e.g. video phone)
- Due to the microwave link, the range of the local loop is only a few hundred meters or 1 to 2 kilometers, requiring more base stations

Bluetooth

The Bluetooth technology has been invented to replace the cabling inerconnecting equipments located within a few meters. According to the Blluetooth standard, three duplex 64 kb/s voice channels or a 1 Mb/s data channel can be implemented. The range is 10 m with a transmitter power of 1 mW. However, there is a possibility to use a 100 m operation mode, but in this case the highest output power is 100 mW.

Frequency Band	Technology
800/900 MHz	Analog cellular, GSM, IS-95, CT-2
1.5 GHz	Tadiran and some other proprietary manufacturers
1.7–2 GHz	DECT, PHS, GSM, and IS-95 variants
2–2.5 GHz	DSC, Tadiran
3.4–3.6 GHz	Nortel, Tadiran, Lucent
10 GHz	Emerging technologies
>10 GHz	MVDS technologies

Figure 2.2.7.3. Summary of available WLL systems

Bluetooth operates in the 2.4 GHz ISM range, similarly to the WLANs. To reduce the effects of interference and multipath propagation, frequency hopping is applied, i.e. the transmit and receive frequencies are synchronously changed 1600 times per second, by using a pseudo random series of frequencies.

Due to the pseudo random frequency hopping, the frequency allocation problemof conventional cellular systems can be avoided. Individual applications can be operated completely independently from each other by using different frequiency hopping series. Frequency hopping requires exact synchronization, requiring the active participation of both communicating devices. The 79 MHz frequency range available for Bluetooth has been divided into 79 parts, each having a widlth of 1 MHz. These are used for the frequency hopping.

In the Bluetooth system, binary frequency modulation is applied, thereby significantly simplifying the transmitter and receiver. Bluetooth systems share the channel according to the time division duplex (TDD) principle. This means that the channel is divided into time segments of 625 μs width, and a specified frequency is assigned to each time segment (this corresponds to a hopping frequency of 1600 hops/s). Consecutive time segments are alternately used for transmission and reception.

The devices communicating over a specified channel form a piconet. Within a piconet, a selected device is designated “master”. This is responsible for traffic control within the piconet, and the other devices participate in the connections as “slaves”. Frequency selection is based on selecting the master device address and clocksignal. Within a piconet, at most seven slaves and one master can be active.

Several overlapping piconets form a scatternet. A specific device can be a member of several piconets, so larger Bluetooth networks can also be established.

2.2.8. Fundamentals of mobile computer techniques

Sándor Imre dr., author

László Pap dr., reviewer

In the followings, we present the fundamentals of the so-called mobile computer techniques which can be regarded as constituting the border line of mobile communication and informatics. From this rapidly developing and far reaching field, a comrehensive survey will be given.

Software radio

In this designation, “radio” refers to a mobile device with properties that can be, by updating the software, continuously expanded and modified according to the technological developments and user requirements. It is fequently abbreviated as SWR or SDR, the latter meaning Software Defined Radio.

SWR systems comprise radios (manual terminals, base stations) having harware elements that can be configured by software programming. This eliminates the need to replace the mobile phone when travelling abroad and using instead of GSM e.g. IS95 CDMA. Only a new software with the CDMA functions has to be loaded into the device. Consequently, a single phone can be used globally, independently from the modulation, bandwidth etc. used in any country.

The development of software radio devices significantly accelerated over the last few years. The intention was to apply directly to the antenna input a high performance A/D converter, and to carry out the consecutive operations on the digitized samples (filtering, conversion, amplification etc.) exclusively by software. This would be the ”ideal” software radio but at present, the technological development does not allow to realize it. According to forcasts, it will be realizible within 8 to 10 years. The functions of this ideal model can now only be approximated, but the approximation will be better and better in thle course of technological developments.

General layout of SWR devices

In this part, the main modules of the manual devices will be outlined. Their realization methods are of utmost importance because their quality is affecting the whole communication system. Figure 2.2.8.1 shows the layout of the SWR terminal.

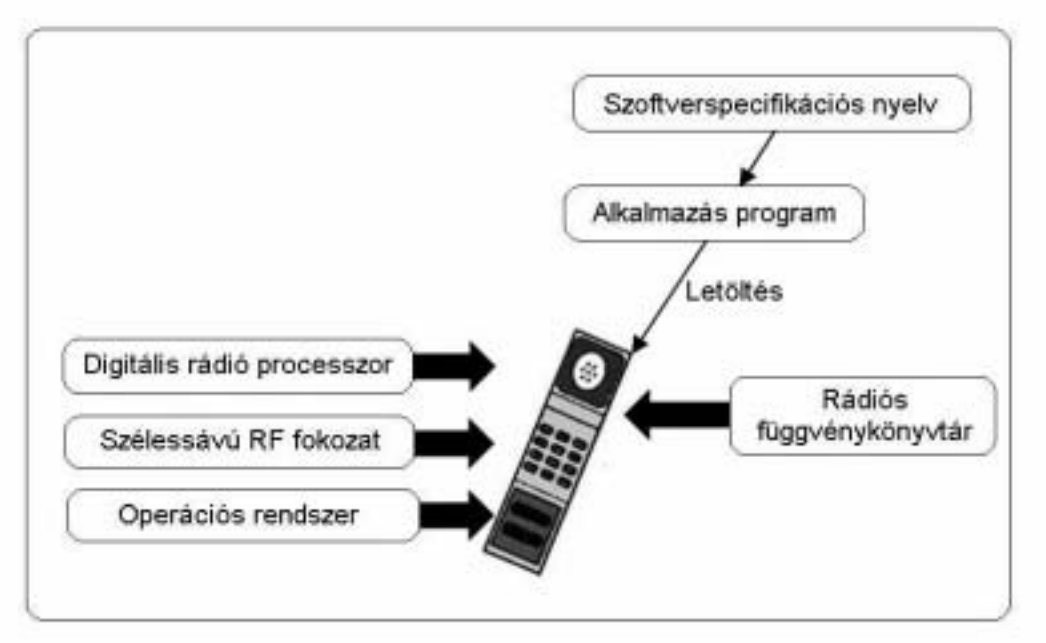


Figure: 2.2.8.1.

Application program, AP

This program is written in a kind of software-specific language, describing the radio architecture and characteristics by utilizing the radio function library. APs are prepared for a specific radio standard, e.g. one for GSM, one for IS95 CDMA, one for UMTS etc. The parameters of the SWR device can easily be reconfigured by re-writing the AP, thereby modifying the use of the function library. APs can be downloaded over the radio channel or even with smart card technology.

Radio Function Library, RFL

This is actually a set of softwares containing basic functions such as the speech condensing algorithm, or the DAC/ADC hardware control programs. These programs may differ even within a single system, or may have individual input parameters, subject to the application in question.

Software Specification Language, SSL

This is some kind of high level language providing the radio characteristics and the exact description for each mobile communication standard. By using this

language, the AP programmer can define the physical layer (transmission power, modulation method, channel encoder, speech encoder etc.).

Digital Radio Processor, DRP

The DRP is a hardware module for general application, comprising typically ADC, DAC, DSP, further programmable logic circuits (FPGA). The modulation method, bandwidth etc. can be determined by modifying the DRP characteristics.

Broadband RF stage

This is a general purpose RF stage for matching the SWR device to systems utilizing different frequency ranges.

Operational system (OS)

Typically, the connection between software and hardwere elements is provided by a JAVA machine. A hardware platform-independent software development is thus possible, simultaneously allowing a competition between the hardware manufacturers

IP mobility

The Internet Protocol has a steadily growing significance in the telecommunication word. Therefore, the methods to harmonize it with the mobility requirements are extremely important. In IPv4, the Mobile IP, introduced to enhance mobility, has the following new network elements:

- *mobile node* (mobile host): This is a terminal that is able to recognize when it is operated far from the host network, and is also capable to start the processes necessary for building up the communication link.
- *home agent* (HA): A computer operating in the home network, typically a router that registers the machines operated far from the network and their actual accessability, and suitably forwards to them the packets arriving at their domestic address.
- *foreign agent* (FA): A router operated in the foreign network that registers the messages arriving from outside, participates in the establishlment of connections with the Home Agent, and possibly in handling the data communication between the Home Agent and the Mobile Host. (The routers having the Home Agent and Foreign Agent function are designated as *Mobility Agents*.)

The process during which the recordings at the Mobility Agents are carried out is called registration. This takes place by message exchanges as illustrated in Figure 2.2.8.2.

From time to time, the Foreign Agent announces its services (1) in its own network. All stations in the subnetwork in question receive these announcements, among others the recently arrived mobile host too. The announcements are carried out by ICMP messages (Internet Control Message Protocol). The message from the Foreign Host is processed by the Mobile Host. In lack of ICMP messages, the Mobile Host sends an addressing message upon which the Foreign Agents operating in the network have to reply. One of these will be selected, and the registration is started.

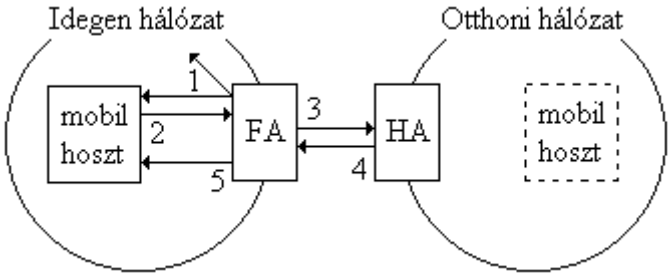


Figure: 2.2.8.2.

The Mobile Host sends a registration request to the Foreign Agent (2). This contains the address of the Home Agent operating in the domestic network of the Mobile Host. This request is processed by the Foreign Agent and is forwarded to the Home Agent (3).

The request is processed by the Home Agent: it will be accepted if the Mobile Host can be served, and if not, then a rejecting answer will be sent to the Foreign Agent (4). Finally, the answer is forwarded by the Foreign Agent to the Mobile Host (5).

The addresses of the prospective Home Agents should be cleared already in the domestic network. The Mobile Host sends a registration request to a well known broadcast address. This will be received by the Home Agents of the subnetwork who will send a rejecting reply (as the host does not exist in the foregin network), and this reply is received by the host. The answering messages contain the addresses of the

sending Home Agents, and these will be stored by the Mobile Host for subsequent applications.

The paths of the so-called tunnels, to be applied in subsequent communications by the registration, are established between the Home Agent and the Mobile Host. The Home Agent has the task to send the packets addressed to the Mobile Host over this tunnel.

The mobile host receives, in addition to its constant IP address in the foreign network, another temporary address; this is the so-called care-of address (COA). The packets addressed to the Mobile Host will be sent to this address by the Home Agent.

This kind of registration is not needed when the host operates in its own network. Following a successful registration, the host can start an IP-based data communication with any machine connected to the Internet. This is illustrated in Figure 2.2.8.3.

After returning to its domestic network, the Mobile Host has to carry out de-registration. This means that it has to send a message telling that the host is at home (it receives the packets sent to its domestic address), and the messages sent to the Home Agent should not be forwarded to the COA. The Mobile Host will again use its own IP address for communication.

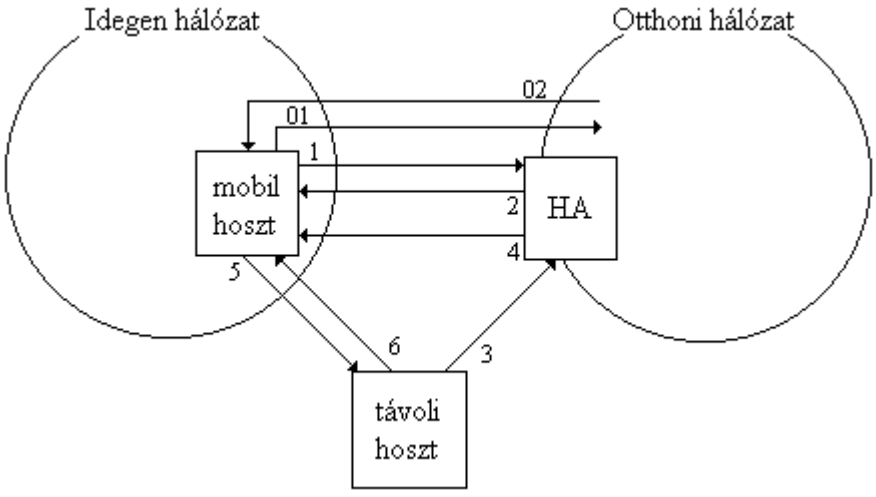


Figure 2.2.8.3

Mobility fundmentals in IPv6

The mobility support provided in IPv6 is different from that in IP inasmuch as no Foreign Agent is needed. Its functions are provided by the network and the Mobile Host itself, by utilizing the new features below.

The Neighbour Discovery protocol is intended to take into account the entities (node machines and routers) within the network. It provides information for the routers and workstations on the accessibilities and addresses of the machines in their environment.

The Address Autoconfiguration is intended to help computers in establishing rapid and simple connection to the network. It also helps equipment connected to the network to obtain the main network parameters: (link local address) and other configuration parameters.

In the case of a *stateless automatic configuration*, there is no need for a separate network server to obtain the wanted data. For a computer to be installed recently, the connection data are obtained from the network by applying the announcement messages explained earlier.

In the case of a *stateful configuration*, the hosts obtain the addresses, configuration information and other parameters from a server. In this case, the server has a data base in which the hosts and the pertaining addresses are registered.

Mobility support in IPv6

As soon as the Mobile Host leaves its domestic network and is connected to another network, a temporary IP address within the address domain of the foreign network will be assigned to it, due to the above mentioned autoconfiguration methods. As soon as the new address is valididated, the host sends a Binding Update to the Home Agent (1) (Figure 2.2.8.3).

Normally, its own Home Agent is known to the Mobile Host, but in some cases, the HA may leave the network (e.g. due to a failure). In this case, the Dynamic Home Agent Discovery (DHAD) procedure takes place: the Mobile Host sends a Binding Update (01) to the anycast number of the Home Agents in the domestoc metwork. At least one Home Agent will respond to the message sent to the anycast

address (02), and sends the list of possible Home Agents. The Mobile Host will then select for itself a Home Agent from this list.

By processing the binding update, the Home Agent will then register the new address of the Mobile Host (Primary Care-of Address), and sends thereof a Binding Acknowledgement (2) to the Mobile Host. The change notice, further the certification will be authenticated by using an IPSec protocol in order to avoid an illegal traffic diversion.

Following the registration of the new IPv6 address, the Home Agent will accept the packets (3), and will forward them to the primary temporary address (4) by using IPv6 embedding.

If the mobile device passes from a foreign network to another foreign network it will again send a Binding Update for his own Home Agent and also to the routers of the abandoned network. These, as temporary Home Agents, will then forward to it the packets arriving into the previous network. After returning to its domestic network, the mobile device informs his Home Agent about its return by sending a suitable Binding Update.

The IPv6 indicates a mobility change also at the nodes communicating with mobile stations. If a remote host is connected to a mobile station then, at first, it sends the packets to the domestic address of the mobile device (3). According to the above procedures, these packets arrive at the mobile station (4). This station detects that the received traffic arrives indirectly, via the Home Agent. Thereupon it sends a Binding Update to the remote host (5) prompting him to store the temporary address and the original domestic address.

During the subsequent communication (6), the remote host will directly communicate with the mobile node, without the Home Agent, if a packet is routed to an address having a stored temporary address. The stations communicating with the mobile device can ask for a Binding Update from the mobile device in the form of a Binding Request.

Agent technology

In the technical literature, some programs having artificial intelligence are called agents. Their most important property is the strong interaction with their

environment by sensing its state, reaching decisions and acting by themselves. In order to reach correct decisions, they need primarily purposes and basic principles to act upon. They also have to measure the success and efficiency of their actions, and get information on their environment.

At present, the agent technology is coming more and more into prominence in the field of monitoring and managing large networks. This results from the fact that in the case of a global telecommunication network, a centralized network supervision can no longer respond to the ensueing events with adequate speed. Some functions have therefore to be distributed, and this is perfectly provided by the agent technology.

The most important agent types are summarized in the following.

- Reflective agents. These merely implement pre-programmed responses to certain events. Their opration is simple and rapid, requiring only the programming of the condition-response pairs. Their advantage is the fact that the “motive”, the frame program, can be prepared independently from the actual problem. However, they can be applied successfully only for a rather limited number of simple tasks.
- Remembering agents. Their operation is similar to that of reflective agents, but they also dispose of an internal status influencing their decision. Due to this internal status, they will “remember”, and act on a certain event individually in different status conditions.
- Deliberative agents. They have not to be acquainted with all possible states, only with their purposes (a wanted status of the environment), and they have to act so as to approximate the object. To do so, they need to measure efficiently the measure of success, allowing them to judge how far they are from reaching their object, and what steps they should undertake efficiently. They are more intelligent than the reflective agents, they can be applied to solve more complicated probems, but their operation is substantially slower due to the many considerations and plannings performed.
- Efficiency oriented agents. These are similar to the deliberative agents, with the difference that they may have several, even contradictory purposes. A decision should also be reached regarding the purpose that has primarily to be served, or regarding the procedure implementing the greatest number of purposes. For the calculation, a well-defined utiliy and cost function is normally applied, weighting the individual purposes according to their importance. The real situation is best modelled by this approximation, but on the other hand, its operation is the most complicated and also the slowest.
- Hybrid Agents. These are mixtures of the precedings inasmuch as reaching the simple decisions with reflective methods, and the more complicated and far-reaching ones by deliberative methods. They are nearly as efficient as the

efficiency oriented agents but in most cases, their speed is higher by many orders of magnitude.

Mobile agents

The mobile agent technology is based on the procedure according to which the task is performed while the codes of the agents, together with the variables specifying their status, are roaming over the network. This approximation allows the agent to adopt itself to the environmental elements such as the software interfaces and services, or the network characteristics. Thus flexible, robust and error-tolerant applications can be found that can be applied in many fields.

To operate the mobile agents, the nodes should be capable to accept and run their codes. The characteristics of the mobile agent systems are summarized in the MASIF (Mobile Agent System Group) standard of the OMG (Object Management Group). Operation of mobile agents is provided by individual solutions suiting the various applications, but general purpose mobile agent platforms are now also in general use, allowing many applications based on mobile agents to be realized relatively simply. The Grasshopper, the IBM Aglets Workbench, the Concordia or the Voyager are systems of this kind. In the following, the Grasshopper mobile agent platform as a general example will be presented.

The Grasshopper system is based on a distributed computing environment, and is completely made up from Java 2 components. The distributed environment itself comprises regions and, within these, nodes called agencies. Each agent needs an agency at which it happens to reside as this provides its running environment. Each agency has permanent agents who are active locally, and mobile agents can also reside in the node. The mobile agents are registered by the region supervision, continuously following the migration and activity of the agents. Via this supevision, a certain agent can exactly be located.

There are certain tasks that all agencies have to perform in order to secure the operation of the agent system. Agents for these basic tasks, constituting the core of the agency, have following functions.

- **Communication**. This service is responsible for managing the links between agents such as message exchanges between agents and migrations, and locating the agents with the aid of the regional supervision.

- **Registration services**. Each agency has to be aware of the agents being runned. This is necessary because of the information to be provided for the external management, and on the other hand, the employed agents have to give account of the registered objects. Also, all agencies are in contact with the region supervision too, requiring continuous updating.
- **Management services**. The management agents have the task to provide the system managers with information, and to perform actions in their name, either due to direct instructions or following the agent's own decisions. Complete control should also be possible. This allows single persons to employ, clear or modify. agents or locations. Further, they should be suitably informed regarding the network, the agencies and the individual agents.
- **Security services**. There is also a need for security arrangements to prevent unauthorized use. Two arrangements have been included in the system. One of these is the classification of the established links and data streams, with the aid of SSL (Secure Socket Layer). The other is the provision of suitable authentication systems, based on the Java security system.
- **Reliability services**. Suitable storage of the agents, yielding restoration of services even in the case of a complete system breakdown.

Translated by Tamás Sárkány dr.

2.2.9. Transmission Methods of Terrestrial and Satellite Broadcast Systems

Imre Kovács dr., author

András Gschwindt dr., reviewer

The terrestrial and satellite audio and video broadcast systems use both analogue and digital modulation. In this chapter, we review briefly the most important transmission methods of the following systems:

- analogue terrestrial audio broadcast systems
- value-added data transmission methods of audio broadcasting
- digital terrestrial audio broadcast systems
- digital satellite audio broadcast systems
- analogue terrestrial video broadcast systems
- analogue satellite video broadcast systems
- analogue terrestrial video broadcast systems
- analogue satellite video broadcast systems

- digital terrestrial video broadcast systems
- digital satellite video broadcast systems
- value-added data transmission methods of video broadcasting

The analogue satellite audio broadcasting is not in the list above, since this system is not developed as a substantive service.

Analogue terrestrial audio broadcasting and value-added data transmission methods

Depending on the band, the analogue terrestrial audio broadcasting use two different modulation. In the so-called long, middle and short wave bands, i.e. from 150 kHz to 30MHz, only one audio channel is transmitted using Double Side Band Amplitude Modulation (AM-DSB). The bandwidth of the modulation signal is usually 4.5 kHz, hence the bandwidth of a single RF (Radio Frequency) channel is 9…10 kHz, and similarly, the distance between two neighboring channel is also 9…10 kHz. Obviously, the taboo frequency allocation is always used which means that the neighbouring radio transmitter can not use neighbouring channels.

The frequency modulating (FM) transmission operates in the VHF (Very High Frequency) band, this method is capable of transmitting one or two audio channels in HIFI quality. In the case of mono transmission only one channel is transmitted by using wide-band FM, where the band of the audio signal is form 40 Hz to 15 kHz. To keep the non-linear distortion below 1%, the required estimated bandwidth of the modulated signal is about 240 kHz.

The two channel stereo audio broadcasting also uses FM, but in this method the modulating signal a special multiplex signal, which is compatible with the mono FM systems. The spectrum of the multiplex signal is shown in figure 2.2.9.1. The spectrum contains the baseband mono signal derived from the left and the right channel (40 Hz – 15 kHz), the stereo signal (S), which is the difference signal of the left and the right channel modulated by using AM-DSB/SC (AM-DSB/Suppressed Carrier), a 19 kHz pilot signal and auxilliary data transmission channels.

Since the 38 kHz carrier of the AM is supressed, the multiplex signal contains a 19 kHz pilot signal, whose frequency is the half of the original carrier, hence this signal is essential to demodulate the AM-DSB/SC signal. The multiplex signal could be generated with the encoder shown in Figure 2.2.9.2. The input signal of right and

left channel is pre-emphasised in the encoder, hence the output right and left signal has to be de-emphasised using an appropriate filter in the decoder. By using pre-emphasis and de-emphasis the demodulated signal-to-noise ratio improved about 12 dB. After the frequency modulation of the multiplex signal the bandwidth is 300 kHz. In general distance of two neighbouring channel is 300 kHz and the taboo frequency allocation is also adopted.

The ancillary data services perform one-way data transmission above the bandwidth of the multiplex signal. The Radio Data System (RDS [2.2.9.1]) is the most common standard for data transmission. The RDS uses 57 kHz carrier to transmit standardized data such as automatic tuning, optimal transmitter selection for mobile receivers, traffic information, personal calls, radio text, etc. The modulation is AM with formed two-phase data and corresponds a two-phase PSK (Phase Shift Keying). The bit-rate is 1187,5 bit/s.

The SWIFT [2.2.9.2] (System for Wireless Infotainment Forwarding and Teledistribution) system uses LMSK (Level-controlled Minimum Shift Keying) modulation with 76 kHz carrier, and the bitrate is 16 kbit/s. The LMSK modulation is a variant of the MSK modulation, where the magnitude of the carrier is controlled by the level of the stereo channel.

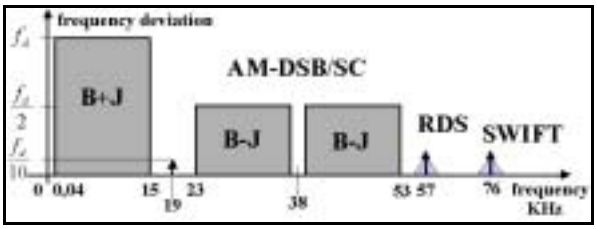


Figure 2.2.9.1 Spectrum of stereo MPX signal

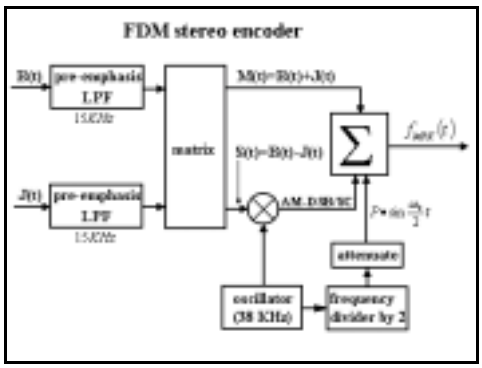


Figure 2.2.9.2. MPX encoder

Terrestrial Digital Audio Broadcast Systems

Recently in Europe there is two systems is used for terrestrial digital audio broadcasting depending on the frequency.

Above 30 MHz the Digital Audio Broadcasting (DAB [2.2.9.3]) is used whose source coding is MPEG audio coding. The DAB permits high bit-rate (about 2.3 Mbit/s) but requires relatively high bandwidth (1.5 MHz). The DAB system uses Coded Orthogonal Frequency Division Multiplex (COFDM) modulation which is supplemented with channel modulation handling the frequency and time invariance of the channel to allow mobile receiving. The most important parts of the channel modulation are convolution error correction, time and frequency interleaving, differential QPSK modulation for each carrier. Four different transmission modes are developed to allow mobile receiving and support Single Frequency Network (SFN) configuration (Table 2.2.9.1.). The bitrates of the audio channels can vary from 64 to 256 kbit/s and must be suitable for the MPEG audio standard.

Mode	Frequency	Application
I	Up to 300 MHz	Terrestrial broadcasting
IV	Up to 600 MHz	Terrestrial broadcasting
II	Up to 1,2 GHz	Satellite and terrestrial broadcasting
III	Up to 2,4 GHz	Satellite and terrestrial broadcasting

Table 2.2.9.1. The four transmission modes of DAB

Below 30 MHz two different and incompatible solutions are under development. The Digital Radio Mundial (DRM [2.2.9.4]) is standardized in these days. The bandwidth of a channel is a multiple of 9 or 10 kHz, hence the bitrate can be adjusted flexible. The source coding is MPEG AAC (Advanced Audio Coding) supplemented with SBR (Spectral Band Replication).

The other system is the In-Band On-Channel Digital Sound Broadcasting (IBOC DSB) which allow the analogue and digital broadcasting side by side, which is useful to change from the analogue to digital broadcasting progressively in small steps. The composite spectrum is shown in figure 2.2.9.3.

Both system uses OFDM modulation for the transmission of the digital signal, the carriers are modulated with QAM, the important parts of the channel modulation: FEC (Forward Error Correction) and long span time-interleaving optimised propagation conditions.

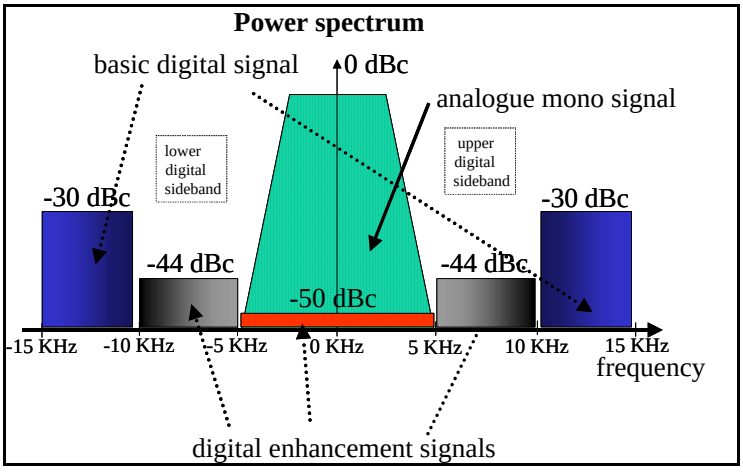


Figure 2.2.9.3. The composite spectrum of the IBOC DSB

Digital Satellite Audio Broadcast Systems

Several Digital satellite audio broadcast systems are developed, but the most important is the DAB which is also implemented to satellite broadcasting. The allowed transmission modes are mode II and mode III (see Table 2.2.9.1.).

Note that several satellite broadcast service providers uses their own audio transmission method, eg the ADR (Astra Digital Radio) which is an MPEG stereo audio stream modulated with DQPSK. The average bitrate of the ADR is 192 kbit/s and the ADR streams are transmitted independently besides the video streams.

Analogue Terrestrial Video Broadcasting

From the beginnings, the analogue video broadcasting uses AM-VSB (Vestigal Side Band) to transmit the luminance and crominance information of the motion pictures. The most current three systems (NTSC, PAL, SECAM) are incompatible with each other. All of them use Frequency Division Multiplexing (FDM), where the main carrier of one channel is modulated by the composite video signal (CVBS: Color Video Blanking Sync), while the audio signal modulates one or two sound sub-carrier. The video transmission is always analogue, the analogue audio transmission is FM, while the digital one uses a variant of QPSK as modulation and NICAM (Near Instantaneously Companded Audio Multiplex) as audio source coding. The detailed specification of the system is in [2.2.9.6]. The compount spectrum of the PAL-CVBS signal is shown in Figure 2.2.9.4.

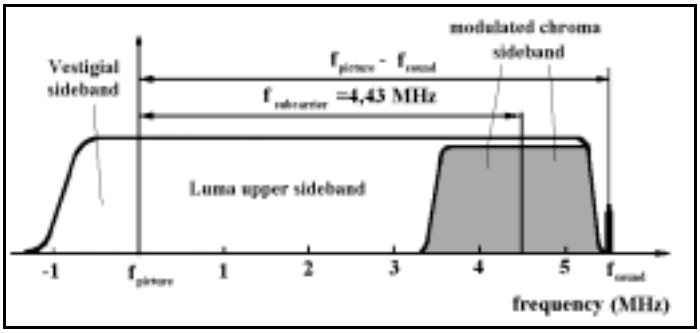


Figure 2.2.9.4 The compozite spectrum of the PAL-CVBS signal

Analogue Satellite Video Broadcasting

The current analogue satellite video broadcast systems are based on three different methods. The American systems are based on NTSC, while the European systems are based on PAL. The third broadcasting system, the MAC/packet (Multiplex Analogue Components [2.2.9.17]) has small significance in these days.

We show the composite spectrum of the PAL-based satellite video broadcasting on Figure 2.2.9.5. The channel allocation is based of FDM, the specification of the composite video signal (luminance and chroma components) is described in [2.2.9.6] hence it is the same as the specification of the terrestrial broadcasting. The composite video signal modulates the main carrier using FM. The audio channels of one video channel are transmitted using additional sound carriers and FM.

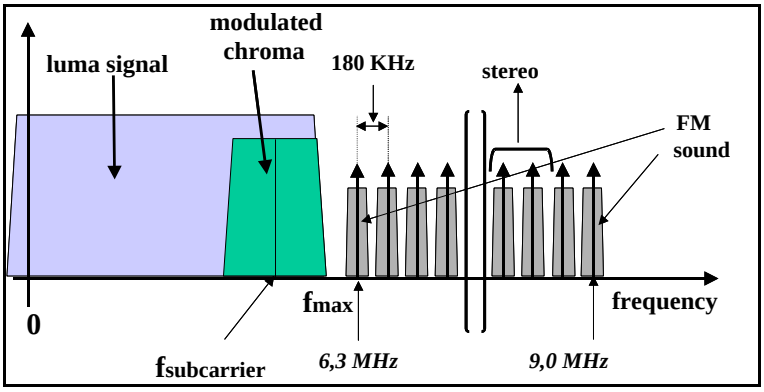


Figure 2.2.9.5. The composite spectrum of PAL-based analogue satellite broadcasting

Digital Terrestrial Video Broadcasting Systems

In Europe, the terrestrial digital broadcasting systems use the DVB-T (Digital Video Broadcasting - Terrestrial) standard. Beside this, the ATSC (Advanced Television System Committee [2.2.9.8]) is used in North America and the ISDB (Integrated Service Digital Broadcasting) is developed in Japan. All of them use the MPEG-2 video standard [2.2.9.9] for video source coding and the MPEG-2 system standard for multiplexing the audio, video and the ancillary data. The audio source coding is different, while DVB-T uses MPEG-1 audio [2.2.9.12], but the ATSC uses AC-3.

In the following, we describe the DVB-T system. The DVB-T standard specifies the modulation and channel coding of the digital terrestrial normal and high definition multi-program broadcasting.

As we mentioned before, the video is encoded with an MPEG-2 video encoder, while the audio is encoded with an MPEG-1 audio encoder. The audio and video components of a program are multiplexed in one steam according to the MPEG-2 system standard. The resulting MPEG-2 TS (Transport Stream) is wound up to the DVB-T channel encoder which performs the following operations:

- transport multiplex adaptation and spectrum spreading
- outer error correction coding (Reed-Solomon 204/188)
- outer convolution interleaving for error correction
- inner convolution error correction coding
- inner time interleaving
- modulation mapping and QAM/QPSK modulation
- creation of the OFDM signal

One of the main goals of the development of the DVB-T was to lower the interference within one channel and between neighbouring channels. Forthermore, the maximal spectral efficiency within the UHF band is also an important requirement, which is achieved by using SFN (Single Frequency Network).

The modulation is multi-carrier orthogonal frequency division multiplex (OFDM). Two operation modes are developed: 2k and 8k. The mode 2k is suitable for small SFN and one transmitter, while the mode 8k is capable of large SFN. The

channel coding uses several different QAM and different inner coding rates, but two-layer hierarchical channel coding and modulation is also allowed.

Digital Satellite Video Broadcasting Systems

In Europe, the terrestrial digital broadcasting systems use the DVB-T (Digital Video Broadcasting - Terrestrial) standard. Beside this, mainly in North America several similar, but not fully DVB-S-compatible transmission methods exist for digital video broadcasting.

In the following, we describe the DVB-S system. The DVB-S standard specifies the modulation and channel coding of the digital terrestrial normal and high definition multi-program broadcasting both in the FSS (Fix Satellite Services) and the BSS (Broadcast Satellite Services) band. The coding system is mainly developed to achieve DTH (Direct To Home) service but it is also applicable for distribution. The main requirement of the development of the transmission system was to create a system that can be adapted to any transponder bandwidth.

The video is encoded with an MPEG-2 video encoder [2.2.9.9], while the audio is encoded with an MPEG-1 audio encoder [2.2.9.12]. The audio and video components of a program are multiplexed in one steam according to the MPEG-2 system standard. The resulting MPEG-2 TS (Transport Stream) drives the DVB-S channel encoder which performs the following operations:

- transport multiplex adaptation and spectrum spreading
- outer error correction coding (Reed-Solomon 204/188)
- outer convolution interleaving for error correction
- inner convolution error correction coding
- base-band spectrum forming
- QPSK modulation

Auxiliary Data Transmission Services in the Video Broadcast Systems

The video broadcasting is capable of several one-way auxiliary data transmission. The most common auxiliary data transmission method is the teletext, which is implemented in both the digital and analogue video broadcast systems. Obviously, transmission of two teletext systems are essentially different. The other auxiliary services are not discussed here.

The teletext system of the analogue video broadcasting has the following properties:

- TDM (Time Division Multiplexing): the data is inserted in the field-blanking interval without disturbing the active picture and audio contents
- binary NRZ coding
- the bit-rate is (6,9375 Mbit/s)
- to control the inter-symbol interference, spectrum forming is performed before insertion

Since the digital video broadcast system uses MPEG-2 TS hence the teletext information can be easily inserted (multiplexed) into the transport stream as a binary stream using sufficient identifiers.

References

[2.2.9.1] Rec. ITU-R BS.643-2 System For Automatic Tuning And Other Applications In FM Radio Receivers For Use With The Pilot/tone System;

[2.2.9.2] ETS 300 751, Radio Broadcast Systems; System for Wireless Infotainment Forwarding and Teledistribution (SWIFT);

[2.2.9.3] ETSI EN 300 401, Radio Broadcast Systems; Digital Audio Broadcasting (DAB) to mobile, portable and fixed receivers;

[2.2.9.4] ITU-R BS. System For Digital Sound Broadcasting In The Broadcasting Bands Below 30 MHz;

[2.2.9.5] NN: SES-ASTRA-ADR, MPM/93-F142B, CD, MPM/93-F115B;

[2.2.9.6] ITU-R Section 11A: Conventional, Enhanced and High-Definition television Systems: Recommendation ITU-R BT.470-4: Television Systems;

[2.2.9.7] ETSI EN 300 744, Digital Video Broadcasting Systems for television, Sound And Data Services; Framing Structure, Channel Coding And Modulation For Digital Terrestrial Television, 1998;

[2.2.9.8] ATSC Standard A/53, Digital Television Standard, 1995;

[2.2.9.9] ISO/IEC 13818-2, Information Technology, Generic Coding Of Moving Pictures And Associated Audio: Video Part, 1994;

[2.2.9.10] ISO/IEC 13818-1, Information Technology, Generic Coding Of Moving Pictures And Associated Audio: System Part, 1994;

[2.2.9.11] ATSC Standard A/52, Digital Audio Compression (AC-3), 1995;

[2.2.9.12] ISO/IEC 11172-3, Information Technology, Coding Of Moving Pictures And Associated Audio For Digital Storage Media at up to 1.5 Mbit/s: Audio part, 1992;

[2.2.9.13] ETSI EN 300 421, Digital Video Broadcasting Systems for television, Sound And Data Services; Framing Structure, Channel Coding And Modulation For 11/12 GHz Satellite Services, 1996;

[2.2.9.14] ETSI EN 300 706, Enhanced Teletext Specification, 1997;

[2.2.9.15] ETSI EN 300 472, Specification For Conveying ITU-R System B Teletext In DVB Bit Stream, 1997;

[2.2.9.16] ISO/IEC 14496-3 Information Technology, Coding of Audio-visual Objects: Audio, 1998;
[2.2.9.17] ETS 300 250, Television Systems, Specification Of The D2-MAC/packet system;

WWW:

www.etsi.org
www.itu.org
www.dvb.org
www.digitag.org
www.drm.org
www.mpeg.org

Abbreviations:

AAC: Advanced Audio Coding
AM-VSB: AM-Vestigal Side Band
ATSC: Advanced Television System Committee
COFDM: Coded Orthogonal Frequency Division Multiplex
CVBS: Color Video Blanking Sync
DAB: Digital Audio Broadcasting
DRM: Digital Radio Mundial
DTH: Direct To Home
DVB-S: Digital Video Broadcasting - Satellite
DVB-T: Digital Video Broadcasting - Terrestrial
IBOC DSB: In-Band On-Channel Digital Sound Broadcasting
ISDB: Integrated Service Digital Broadcasting
LMSK: Level-controlled Minimum Shift Keying
NICAM: Near Instantaneously Companded Audio Multiplex
RDS: Radio Data System
SFN: Single Frequency Network
SWIFT: System for Wireless Infotainment Forwarding and Teledistribution

Chapters 2.2.1., 2.2.2., 2.2.4., 2.2.5., 2.2.7., 2.2.8. translated by Tamás Sárkány dr.

3. Switching technology

The high quality switches already implemented in telecommunication networks will play an important role in carrying the traffic for a long time. On the other hand we face revolutionary changes in the quickly spreading area of mobile and multimedia communications and the Internet. All these new technologies and methods are built heavily on communication protocols and software.

This section gives an overview about the world of exchanges in a retrospective way but at the same time it brings the reader into the world of communication switches. (Section 3.1, 3.2.)

The old exchanges contained the followings:

- The exchange board,
- The intelligence needed for operation,
- The signal routes and the
- Signals controlling the exchange.

The routers of the Internet and the modern exchanges have the same features basically, nevertheless some further functions as accounting, statistics about the operation and traffic and the capabilities in automatic self-reparation are in most cases more sophisticated.

A separate section is devoted to the signalling system of digital switching, summarizing the system's architecture and protocols of mobile communications.

Traffic issues are new big challenges in modern switching technologies. In Section 3.3 the reader can learn about planning optimal traffic management.

New solutions and services can be realized on the existing networks with modern communication protocols. Section 3.9 leads the reader into the design process of the protocols. Formal description languages and techniques that are widely accepted and used in the telecommunication world play an especially important role in the protocol design process. The advanced protocol engineering methods make it possible to develop large software and protocols but conformance to the specifications should be always tested. Section 3.10 describes the procedure of conformance testing.

IP technologies are main building blocks of communication networks. Sections 3.6 and 3.8 deal with IP technologies and the services it can offer. ATM technology is also widely used in our networks, as described in Section 3.7.

Big efforts are put into providing new and guaranteed quality services on the modern networks nowadays. The following sections give an overview on switching and signalling techniques, IP, ATM and GSM systems as well as on the discussion the world of communication protocols in details.

Gyula Csopaki, Sarolta Dibuz, Editors of the Chapter

3.1. Overview of Switching Technology

Gyula Csopaki, author

György Lajtha, dr., reviewer

This paper gives an overview of the development of the space-division subscriber switching exchanges of the Public Switched Telephone Networks (PSTN). According to today's classification these exchanges were based on the principle of circuit switching. Circuit switching was the only dominant principle in switching technology through a period of more than 100 years. This technique was used in recent systems like telex, narrow-band ISDN (Integrated Services Digital Network), mobile telecommunications and what is more in the first data networks as well. For example, the data and telex centre, controlled by stored program and set up in Budapest in the early 1980s (NEDIX), also used this technique.

Main Functions of Switching Exchanges

Generic block diagram of a switching exchange is described in Figure 3.1.1. The schematic descriptions generally show the switching-network being two-sided: on one side the circuits to be connected (channels), briefly the inputs can be seen; on the other side the circuits appropriate for the demand, briefly the outputs can be found. Most of the exchanges in practice are structured according to this arrangement. Still, there are some structures, where both circuit groups are to be found on one side. These sorts of implementations are called folded back exchanges.

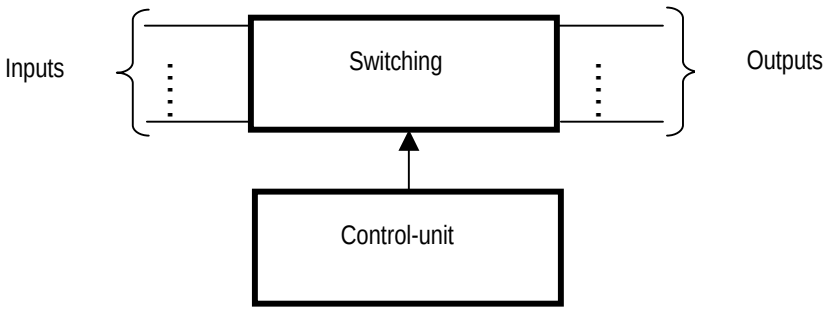


Figure 3.1.1: Generic block diagram of a Switching exchange

The switchboard connects the inputs and outputs temporarily. Switching can be space-division, frequency-division or time-division, i.e. the simultaneous connections are separated in space, frequency or time domain, in the switchboard.

The main function of the control-unit is to set up the demanded connection in the switchboard. The calling party gives (subscriber's terminal equipment or another centre) gives the address of the place of destination to the control-unit. On the basis of this addressing the control-unit determines the appropriate output and chooses a free interconnection route among the routes for connecting the two end points, allocates this and sets up the connection using the appropriate switching units. The control-unit also informs the output. (In case of analogue subscriber's line the controller sends a ringing signal through the line. In case of circuit between two exchanges the controller sends all pieces of information to the other centre on the basis of which the latter can connect to the destination address.)

Another important task of the exchange beyond the switching functions is to receive, evaluate and send all the necessary information about connection management. This is carried out by signalling systems. Signalling systems can be divided into two big groups: subscriber signalling systems and inter-exchange signalling systems. Subscriber signalling systems are simpler than the signalling systems among centres, partly because the local centre communicates to the users with signal tones in many cases and the users themselves evaluate these signals, on the other hand there are signals for operation among exchanges.

Space-division central exchanges

The first switching was set up with 21 subscribers almost 2 years after invention of telephone (on 28 January, 1878) in New Haven (Connecticut, USA) (1). Hungary tried to keep pace with this development; its first central exchange with 50 subscribers was set up in Budapest, on 1 May, 1881. Shortly after this several centres were set up in Budapest and in bigger towns in the country (3).

Types of the space-division central exchanges can be grouped into three categories: manual switching exchanges, electro-mechanical switching exchanges and stored program controlled exchanges.

The two main units of the generic block diagram of the **manual switching exchanges** were separated in reality as well: on one hand there is the switchboard and on the other hand there is the operator. In the developed models the subscriber's lines ended in the jacks mounted in the vertical plane of the switchboard and next to them there was the announcing equipment. The operator paired jacks with cord terminated in both side in jack connecting the subscriber's lines. In the beginning the feeding of microphones of the subscriber's set was from a local battery (LB). The calling subscriber signalled his calling intention to the centre with an inductor in the subscriber's set; and in the beginning even the operator gave a ring with an inductor to the party called.

In the 1890s, as a result of technical development, feeding from local battery is gradually replaced by feeding from central battery (CB) and the feeding current automatically transmitted the subscriber's signals (position of the handset) to the centre. While hook the handset a contact pair closes the subscriber loop with direct current and this current is monitored by a relay, which switches on a lamp next to the jack.

Automatic cord circuits of the CB exchanges replaced part of the work of operators so this way the operators could handle considerably more calls.

The second generation of switching technology are the **electro-mechanical switching exchanges**. Among the several systems only some significant centres and the system applied in the domestic network are described here.

After the centres were set up many people were, especially in the USA, dealing with replacement of operators by machines. Almon B. Strowger patented the first solution that could be used practically as well in 1891 (1). The **STROWGER** switching device consisted of 100 arc terminals on 10 levels, 10 arc terminals on each, arranged in half circle, and a frictional wiper, which could be contacted to the arc terminals. Subscribers directly controlled the frictional wiper to the appropriate level and then to the appropriate position. A lifting device and a rotating magnet moved the wiper. The magnets changed the position of the wiper gradually by being energized (lifting and rotating). This operation method is called step-by-step (SxS) method. Putting another selecting stage before the intercommunication plant there will be a choice among 10 line selectors so this way the capacity of the centre can be

upgraded up to 1000 lines. The upper limit value of a local centre in the Strowger system is 10000 lines, which can be reached by applying other pre-selector stages.

The number of the pre-selectors has significantly been decreased as a concentration stage was inserted among the subscriber's lines and the first pre-selector stage. In this way the generally used switchboard structure of the local centres stepping mechanism assembly, as described in Figure 3.1.2, was developed.

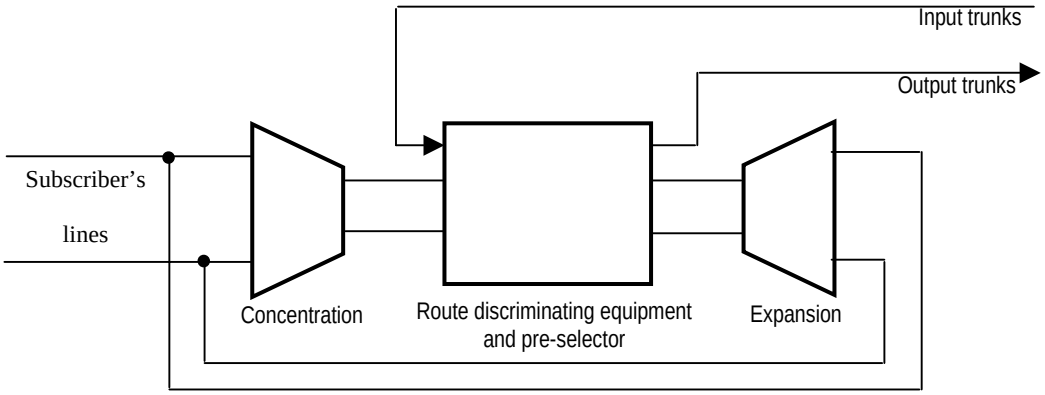


Figure 3.1.2 Switchboard structure of local centres

The separate concentration and expansion stage were coupled in the centres where choosing the link and controlling the switching units are totally separated from the switching unit, like in the cross-bar and the stored program control exchanges. The coupled stage is called the subscriber stage.

The SxS systems have spread all over the world under different trade marks (Strowger, Siemens, 'director' in Britain etc.). Apart from the British 'director' system all the systems used **direct control**. Advantage of direct control lies in its simplicity but subscribers' call number, dependent on the capacity of the centre and its position in the network hierarchy, was of different length (free numbering). It was rather inconvenient for subscribers when the trunk call appeared. (From a different place you might have to dial a different number to reach a subscriber.)

Another big family of rotary brush switching devices is based on the principle of register controlled **indirect control** and it is also called motor driven systems. Switching device brushes are moved by a common axis.

A register receives and stores the information of the calling subscriber and then controls the build-up of the call. The impulse control of selection enables using other systems than the decimal system.

Technical arguments (producing a simpler switching device) and telephone traffic arguments are also for the impulse control of selection. It has been observed and later proved by Erlang's B formula that in the busy hours the usage of circuits (i.e. their performance throughput) increases if the number of servers in one group increases. For example, next to a blocking probability of B = 0.005 the single circuit availability of 10 circuit is 0.4 erlang, of 20 circuit is 0.55 erlang and of 30 bunch is 0.63 erlang, respectively.

Among rotary brush switching devices with impulse control of selection the best-known is the family of ITT Rotary (the number '7' designated the system family, letters standing next to '7' designated members of the family. Better-known centres of these are: 7A in big cities and 7DU in little towns) and LM Ericsson's AGF centre. In systems 7A and AGF control of selecting stages was carried out by revertive impulses, i.e. switching devices sent back impulses to the register through speech wires while moving.

In a recent version of system 7A, in system 7A-2, the capacity of selector was 300 arc terminals on 10 levels, 30 arc terminals on each. The choice here is also bi-directional but lifting up of the contact brush is substituted by brush activation. (The brush carriage contains 10 brush sets as for the 10 levels.)

In line selection function the selector connects 400 stations including 200 stations and 200 party-line stations. Party-line station lines common arc terminals are connected to the central arcing terminal block of 100. The primary aim of using party-line stations is to increase the numbers of subscriber's lines. A solution to this, developed by Hungarian engineers, resulted in expanding the capacity of the switching exchange (3). Using two pre-selectors the capacity of the switching exchange is 40,000 subscribers. A one-motion, search device was used in the concentration stage and in the register switching stage.

Another member of the Rotary family was the 7DU system. In this system in every switching stage a one-motion switching device with 100 arc terminals was applied (its production is simpler) and the register controlled the switching devices by

forward impulses. Digits sent from the register were not received by the switching devices but the control circuits, which were common for more than one (e.g. 10) switching device. On the basis of the received digit(s) the control circuit electronically designated the arc terminal position (line selection) or the arc terminal group (pre-selection) position that the brush of the switching device had to be controlled to. Afterwards the switching device searched the selected position. The intertwined arc terminal block enabled creation of circuits in the pre-selection stages other than decimal circuits, if needed on the basis of traffic conditions.

Switching device of **AGF system** contained 500 arc terminals where grouping was 25x20. Arc terminals and multiplication (wiring of same arc terminals of devices within one group) are replaced by rigid enough, blank wires extended on frames, which resulted in considerable savings both in material and work. Multiplication is entirely separated from the switching device and consists of 20 frames each. Frames are set up radiately and vertically along a circular arc shorter than a semi-circumference.

In the flat switching device the brush assembly is mounted on a rotating disc, by the turning of which the brush can be set up to the appropriate line and then moved radiately. After terminating the connection the brush assembly will move into the opposite direction, which was achieved by reversing the rotation sense of the connecting, constant-rotating axis. Switching devices are plugged-in device units thus the number of devices on one frame can be changed flexibly (e.g. increase) and the devices being out-of-order can be replaced easily. The same type of device is used in the pre-selector, line selector and call concentration stage in the switching exchange. Call concentration stage means that the subscriber’s lines belong to only one multiplication, and line selectors and call finders can be found on the same frame. Robustness of the system was proved in the 1950s when after transplanting the centre from Miskolc, set up in 1930s, to Eger, the system operated well.

The spreading of trunk circuits and the increase in the number of subscribers made the weaknesses of rotary brush switching devices clear: slow operation, noisiness and big maintenance demand (5 hours / year / subscriber). To decrease these parameters the relay is the most appropriate tool in electro-mechanical technology but the relay switch matrices are expensive because of the necessary magnets in each crossover points.

In Figure 3.1.3 there are 4x4 switch matrices described with one-line relay contact symbol without indicating the operating magnets. (The interconnecting unit in the intersection point of coordinates is called the crossover point.) Interconnection of a given input and output means operating the crossover point found in the intersection point of them (closing) and holding it closed meanwhile (interconnection memory). If number of verticals is denoted by ‘n’ and number of horizontals is denoted by ‘m’ then the unique control of crossover points requires ‘n x m’ magnets.

Through coincidental selection the number of magnets can be decreased to a number of ‘n + m’, if the effect of magnets is transmitted to crossover points by an appropriate method, for example low-rate turning of bars. The **crossbar switching device** has been named after these bars. (The crossbar switching device was already created by Swedish Betulander in 1919.) The column magnet with the bar and the contacts mounted along the vertical plane are manufactured as one mechanical unit. The fixed contacts are multiplied in form of plate or wire. The unit is called a bridge, which is similar to a relay with ‘m’ contact spring nests. The horizontal magnet, or marking magnet in other words, has the task to determine which crossover point of the bridge should close. Contacts are made of precious

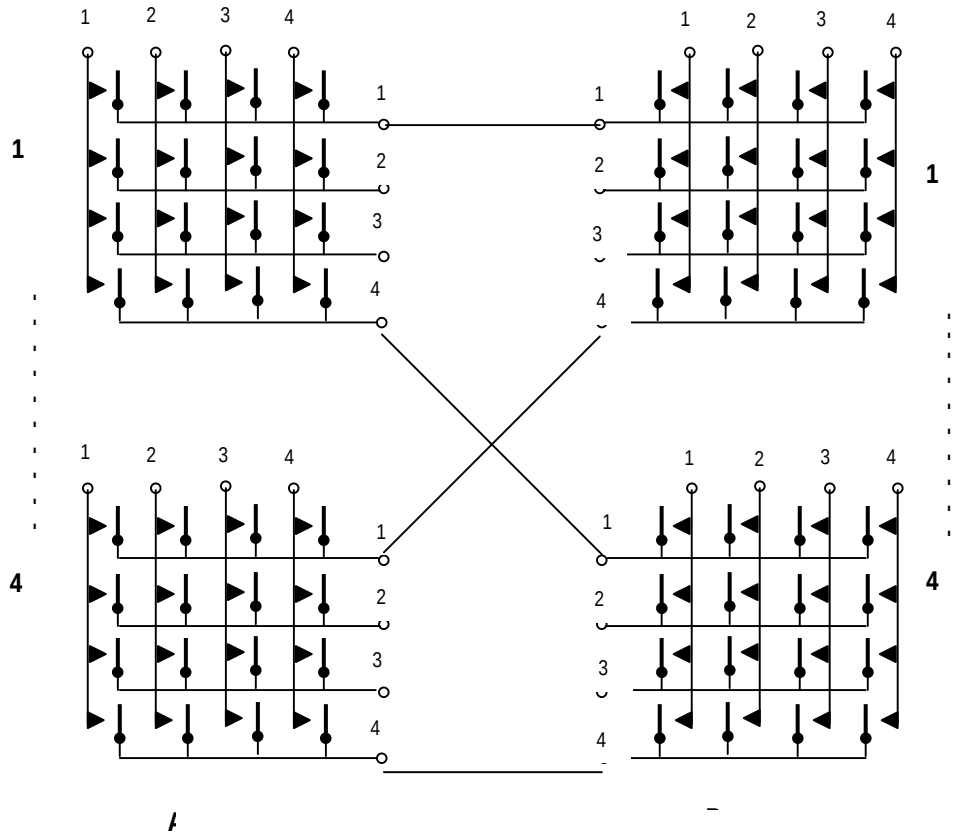


Figure 3.1.3: Double-stage link interconnection

metal ensuring good contact. Economic aspects also demanded applying bridges with a small number of crossover points (10 or 20). Bridges mounted on one common mechanical frame with the attached marking magnets and bars are called together a crossbar switching device. Through multiplication of horizontal outputs of the bridge a matrix of arbitrary size can be produced, the common mechanical frame and designation does not affect the size of the matrix.

Switches with several outputs can be composed by cascade connecting of switching matrices with a small number of crossover points. A double-stage, 16x16 matrix, composed of 4x4 matrices, link interconnection, is described in Figure 3.1.3. The number of built-in crossover points is 128. A 16x16 switching matrix realizing a switch with the same capacity needs 256 crossover points. Link interconnection decreases the number of crossover points needed in the switchboard, however, there is a counterpart: choice of a free output depends on the condition that the link leading there is also free.

There is a possibility to compose 3-, 4- or more-stage interconnections as well. (In order to increase the number of outputs and inputs or to increase the interconnection possibilities – routes.) Link interconnection was introduced in the crossbar systems and since then every switchboard with high capacity (even the digital ones) is produced on the basis of this principle.

Switching exchanges based on the crossbar system appeared in the USA in the late 1930s. Developed version of the AT&T crossbar local exchange was designated **No. 5 Crossbar system** (4, Chapter 8), and its schematic block diagram is described in Figure 3.1.4.

Switchboards consist of line links and trunk links. Both modules are double-stage link interconnections, the switching device contains 20 bridges with 10 crossover points.

Control functions have been split more. Complicated interconnection control functions have been taken out from the registers with relative long holding time and these functions are executed by the markers. Holding time of these markers is shorter thus there is a need for fewer of them than those of the registers. Still, they are very complex. Designing highly-complex marker circuits required the application

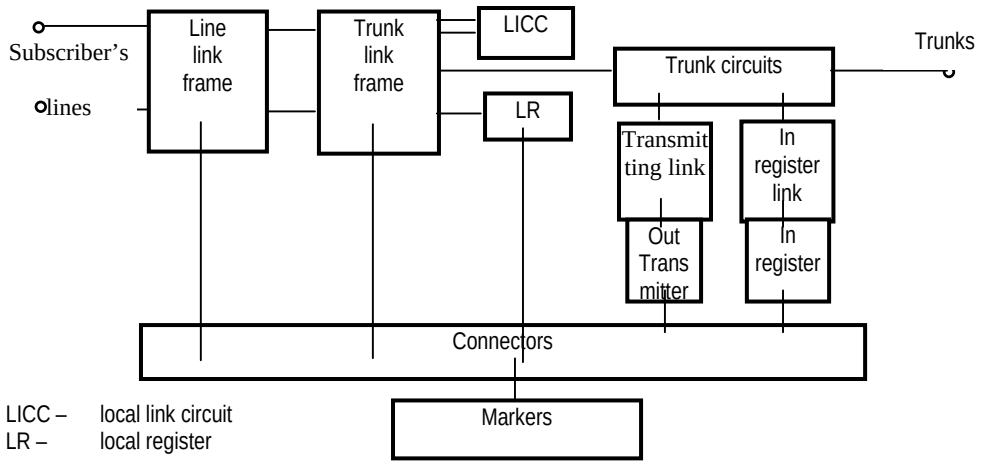


Figure 3.1.4: Simplified block diagram of No. 5 Crossbar

of scientific methods. This was the time when the principle of designing logical circuits was born.

The principle of **group operation** was applied at controlling the switchboard. Searching and then setting up a free interconnection route is simultaneous in both modules. There is a need for fewer interconnection routes in the common-controlled switchboard than in the stage-by-stage-controlled switchboard. Markers connect to the modules, registers and markers through tap switch connecting circuits.

The first crossbar centre was set up by LM Ericsson in Europe in the early 1950s (1, Chapter VIII). Block diagram of **ARF 102** centre, a centre from the AR^4 crossbar family for big cities, is shown in Figure 3.1.5. The Ericsson crossbar device contains 10 bridges with 20 crossover points, 10 marker magnets and 2 selector magnets. For a simpler overview the bridge in the diagrams was described with the same symbol as in the rotary brush switching device. The bridge is equivalent to the brush, the crossover points are equivalent to the arc terminals, so this way connection of links among stages can easily be traced.

The ARF system is controlled step-by-step, which results in some more interconnection routes but the marker circuits are considerably simplified. The subscriber's stage concentrates the output traffic with two stages (SLA – SLB), and expands the terminal traffic through four stages. The number of outputs of the basic double-stage pre-selector module can be expanded more by adding a third stage, according to the capacity and traffic demands.

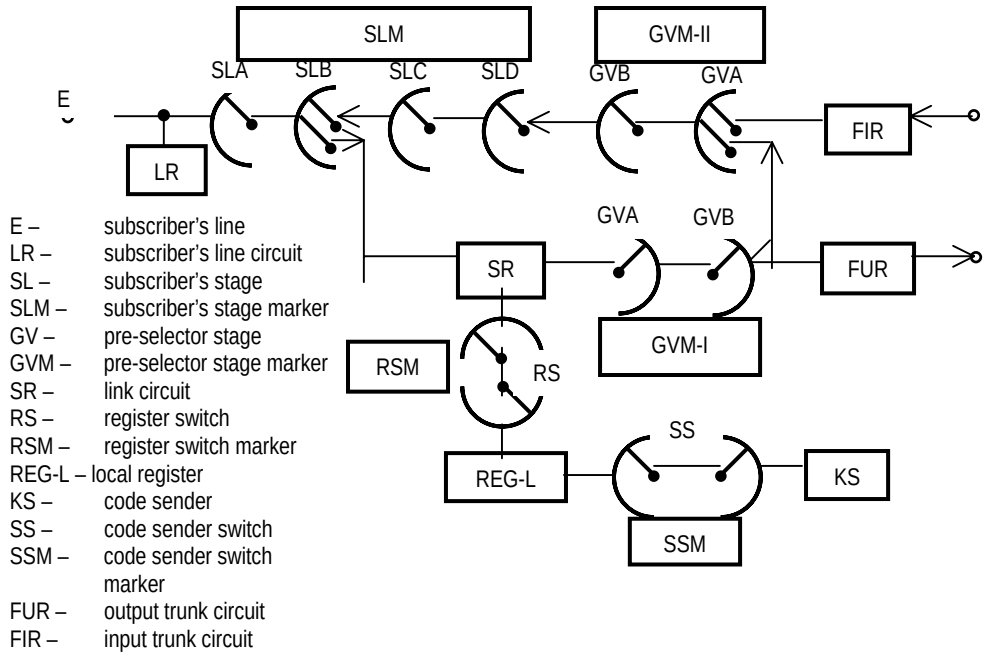


Figure 3.1.5: Block diagram of ARF 102 centre

Because of economy and co-operation requirement of centres with different signalling protocols the forwarding function, which forwards information of choice, was taken out from the register even in this system. This forwarding function is carried out by the code-senders. The inter-register signalling protocol of AR system is the R2 MFC signalling protocol. The code-sender may change the code-sender (signalling protocol) even during the establishment of a call with the help of a switching stage. Segregation of code-sender from the register is even more advantageous because alterations, needed while operating, have to be made at fewer places.

Several other companies designed crossbar system centres. These centres were not of great significance compared to the previous ones. Although the ECR family was not mass applied but this electronically controlled rural centre system, developed in the early 1960s by the Hungarian industry, is worthy of mention (5, p. 109).

The first **stored program controlled centre**, the **No. 1 ESS** (Electronic Switching System) centre (5, Chapter 11), developed in Bell Laboratories, was set up in May 1965, as a result of world-wide intensive research (2, Chapters II And V). The research aimed at electrification of central exchanges. More increase in hardware demand, i.e. more economically produced and operated centres was expected from

higher speed of more reliable electronics, on the basis of the time-honoured system of, according to today's terminology, wired-program controlled crossbar centres. Entire electrification could not be realized with the components available at that time (high-vacuum tubes, gas-filled cold-cathode tubes, early transistors and diodes). For example a quite new device had to be developed for the electronic crossover point switchboard, which was not economical because in those days replacement of already existing centres would have involved a great number of devices. Moreover, the big space-division electronic switchboard did not even meet the requirements of transmission technique (interference).

In Bell Laboratories they became conscious of the fact that requirements of the manufacturer and operator (economic efficiency, expanding and introducing services) could be met by stored program controlled centres. Control of calls can be described as a sequence of simple logical and arithmetic operations, which can be stored in and retrieved from memory. To deploy the processor's processing time well, the search and select operations are not to be executed in the real switchboard but on its projected mirror image in the memory (map-in-memory). Thus the frequent and slow input / output operations are neglected. Schematic block diagram of the system, which is to be seen in Figure 3.1.6, practically reflects division in Figure 3.1.1 (part 1 and 3), completed by communication means between the two parts (part 2).

The switchboard was produced from matrices with ferreed (ferrite + reed) crossover points. These crossover points are controlled by current impulse and are magnet holding, metal-to-metal contacts.

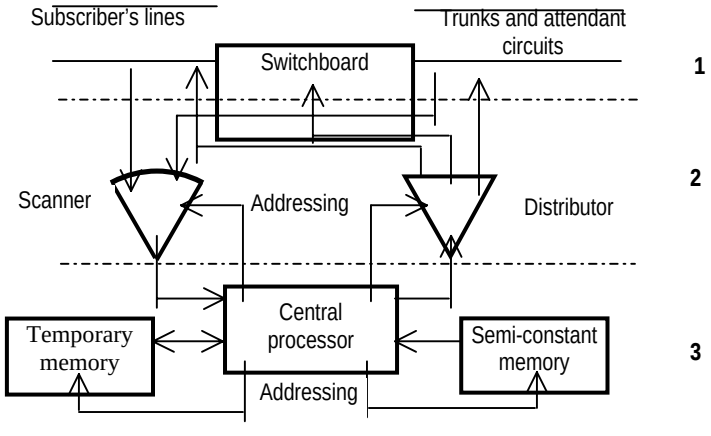


Figure 3.1.6: Schematic block diagram of No.1 ESS

The central processor was made from discrete semi-conductors so it was very expensive. Thus every task of information evaluation, processing and storing was integrated in the processor (centralised control), circuits connecting to the switchboard contained only minimal hardware. Data related to the processing of calls was stored in ferrite memory, programs and semi-constant (centre-dependent) data was stored in twistor (non-destructive) memory.

Safety operation of the centre (a maximum of 2 hours drop-out in 40 years) was ensured by duplication of the control system, i.e. processor and memories. The controls operated in synchronous duplex mode of operation: both processors process the same call but only one, the active one, controls the telephone peripherals. Switching between the controls, due to a fault or any kind of reason (e.g. operator's command), can be carried out without losing a call.

The scanner scans the state of the telephone peripherals periodically, stores these data in its memory until the next period as to determine change of state marking the service demand. (Change in state is always an event.) In order to meet the real time operation requirement for call detection, the subscribers' lines were scanned in an average time period of 100 ms, while the lines in dialling phase were scanned strictly in a time period of 10 ms. This frequent but simple routine took up significant processing time of the processor. Therefore line capacity of the centre was much below the planned one. In order to increase call-processing capacity, in a later version of the device the routines were performed by a pre-processing processor, the Signal Processor (also duplicated) and gave the events over to the central processor to be processed later.

Exploiting results of the development in technology (processor 1A, remreed etc.), size, consumption, producing and operating expenses of the system have significantly decreased. In the second half of 1960s several companies, producing switching exchanges, developed stored-program controlled centres. These were practically later versions of No. 1. ESS. Also a metal-to-metal contact was used there, but the current holding version of reed relay, or special switches holding without current (so as to decrease the size of magnets) and operating on the basis of crossbar principle. The French CIT Alcatel company applied the load sharing mode of operation as a mode of operation for the duplicated processors. Double-stage control replaces the centralised control after fall in price of the microprocessor. The

hardware-close functions, the routine operations are placed in controls of the telephone peripherals. These peripheral controls are considered as distributed Signal Processor of No. 1 ESS. These systems were designed in a way that results of the development in technology (e.g. time-division interconnection) can be built in without any basic changes in the system (see centres described in Chapter 3.2).

Time-division switching exchanges

Electrification of switching exchanges has already been parallel to the development of space- and time-division switchboard from the beginnings. In the case of time-division switching exchanges there are no interference problems and through multiple utilization of elements it promised better switchboards. The experimental systems, apart from one, applied the PAM technique. Two systems are worthy of mention: one is the British Highgate Wood centre, set to test operation in the early 1960s (2, Chapter II-5), and which fell through because of interference in PAM systems. This failure was significant because the British Post decided to skip crossbar systems and realize the needed network modernization with time-division centres. The other is the subsystem of Bell Laboratories, named No. 101, set up simultaneously with the No. 1 ESS centre and used its technology to a great extent (2, Chapter V-1).

Simultaneously with spreading of PCM transmission systems, some people in Bell Laboratories experimented on directly connect channels of digital transmission routes. The experimental system (ESSEX) (2, VIII-2) is considered to be the forerunner of first generation digital centres (here the concentration was carried out by space-division stage). The experimental model justified the practicability of digital time-division interconnection but it was too expensive because of lack of the suitable components.

Cheap semi-conductor memories made the digital interconnection economical, first in transit centres where a connection has to be set up among already digitised transmission routes or among analogue group of trunks that can economically be digitised with group codecs. The first digital transit centre, the No. 4 ESS centre of Bell Laboratories, was set up in 1976. The spectacular development of VLSI technology in the 1980s made the realization of unique codecs possible, as a result of which digital technology reaches to the connection of subscriber’s line, i.e. IDN

(integrated connection and transmission) is established, then through digitising the last stage (subscriber's access) ISDN is established.

[1] Parameters of ringing signal used in recent centres (voltage and frequency) derive from here.

[2] The Hungarian central exchange was automatized with rotary system centres. The first 7A centre was set up in Krisztina building in 1928.

[3] For example there is only one arc terminal is efficiently busy at an interconnection made by rotary brush switching device of 'N' arc terminals while there are 'N-1' arc terminals being busy in vain. It was acceptable at arc terminals made of phosphorus bronze.

[4] Hungary bought the licence of the system in 1968.

References

[3.1.1] Robert J. Chapuis, 100 Years of Telephone Switching (1878-1978), Part 1: Manual and Electromechanical Switching (1878-1960's), North-Holland Publishing Company, 1982.

[3.1.2] Robert J. Chapuis, Amos E. Joel, Jr., Electronics, Computers and Telephone Switching. A book of technological history as Volume 2: 1960-1985 of "100 Years of Telephone Switching", North-Holland Publishing Company, 1990.

[3.1.3] Postamérnöki szolgálat 50 éve, 1887-1937, 2. kiadás az eredeti alapján, Budapest, 1990, Távközlési Könyvkiadó.

[3.1.4] Richard A. Thomson, Telephone Switching Systems, Artech House, 2000.

[3.1.5] A Magyar Híradástechnika Évszázada, MTESZ Házinyomda.

3.2. Digital exchanges

Péter Seres, author

Béla Frajka, reviewer

3.2.1. Introduction

Digital exchanges, as classified by their network traffic function, may perform either as subscriber exchanges or as transit exchanges (trunk exchanges). However, in many cases, they act as *combined exchanges* having diverse traffic tasks within a single exchange. This combined application is made viable primarily by the applied *software elements*. It should be noted that in the Hungarian network, the transit exchange function might cover any call processing applied in a primary, secondary or tandem exchange. In the following Sections of this Chapter, the architecture and main system parameters of up-to-date digital exchanges will be discussed.

3.2.2. Architecture

Figure 3.2.1 shows the functional layout of a digital exchange. The functional units in this Figure may be realized either by hardware or by software, but “mixed” units can also be applied, depending on the requirements. Our investigations will be strictly confined to the main tasks that have to be performed by the units designated by rectangles. (Shading is used only to facilitate the distinction between units.)

The investigated digital exchange has a *combined function* as it can be accessed not only by subscribers but also by other exchanges (by both subscriber exchanges and transit exchanges). To this end, the main functional units of the exchange are the following:

- subscriber stages for interfacing and traffic concentration of the subscriber stages,
- trunk processing stages for processing the inter-exchange signalling,
- digital switching network for providing the switching functions,
- stored program control governing the exchange operation.

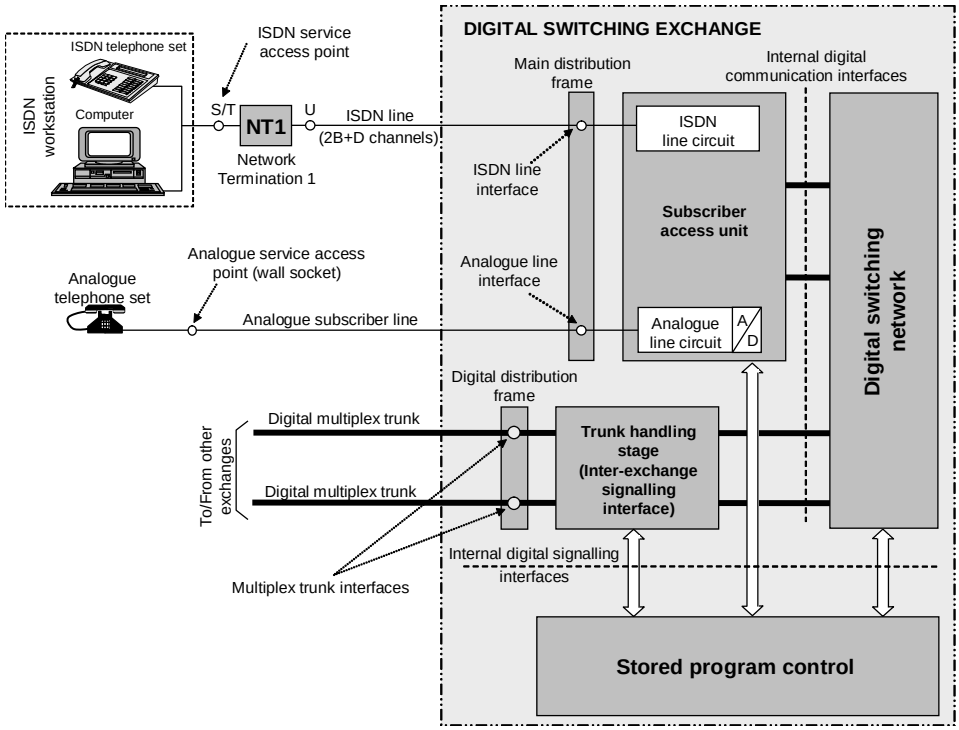


Figure 3.2.1. Functional layout of a digital exchange

The exchange control should be able to handle following traffic situations, even in any simultaneous combination:

- handling of local and outgoing traffic, a subscriber exchange function,
- handling of incoming traffic terminated on subscriber lines, likewise a subscriber exchange function,
- handling of incoming transit traffic, a transit exchange function.

3.2.3. Main switching stages

The stages covered in this Section are intended for *routing* a selected connection.

a) Subscriber stages (digital/ISDN and analogue subscriber lines)

Figure 3.2.2 shows the functional layout of the Subscriber Access Unit. To meet the requirements of two kinds of subscriber accesses we have two types of *subscriber line circuits* (ISDN and analogue subscriber line), and two kinds of line boards. The number of line circuits on a board is determined by the exchange type, and also by the manufacturing technology. Typically, boards comprising 4, 8 or 16 line circuits are applied.

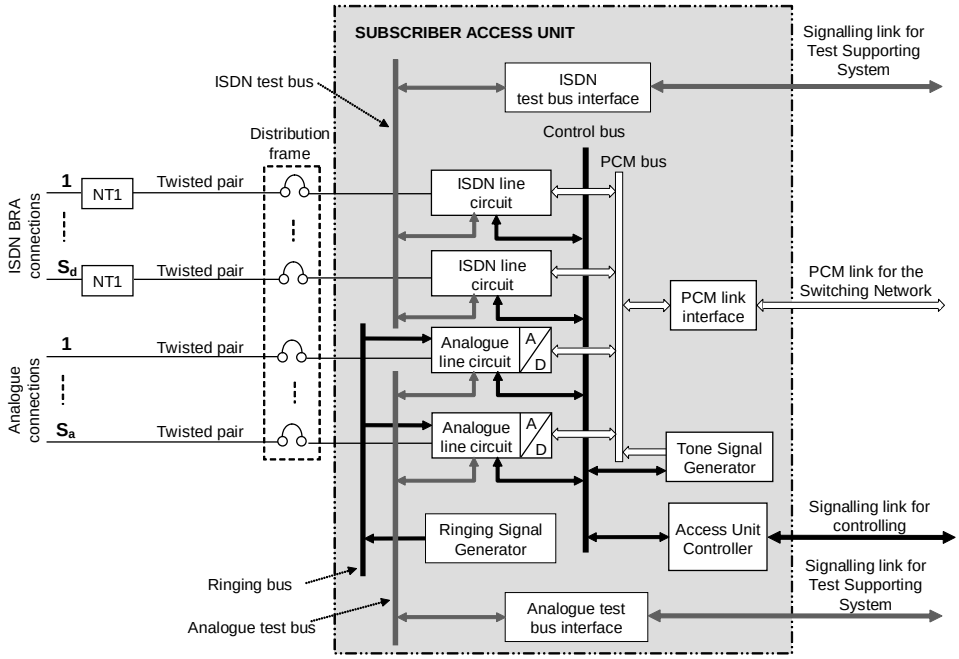


Figure 3.2.2. Functional layout of a subscriber access unit

Data transfer between the ISDN terminal equipment connected to the *digital* service access points and the subscriber stage takes place via the *D-channel* of the connection, according to the rules of the DSS1 subscriber signalling system.

Communication between the terminal equipment connected to the *analogue* service access points and the subscriber stage takes place via DC and voice frequency signalling transmitted over the subscriber line. To meet the required environmental conditions, the analogue line circuits have to realize the so-called BORSCHT functions as presented in the following:

- Battery feed (B),
- Over-voltage protection (O),
- Ringing (R),
- Supervision (S),
- Coding (A-D conversion and filtering, C),
- Hybrid (A-D conversion, H)
- Testing (T).

A digital signalling tone generator, common to all subscribers of the unit (both ISDN and analogue subscribers) is applied for supplying suitable information tones required by the traffic situation.

A separate ringing generator is applied in the Subscriber Access Unit for ringing the called subscribers having analogue lines. The frequency and AC voltage of the signalling generator are chosen to actuate both up-to-date electro-acoustical tone ringers and traditional electro-mechanical bells.

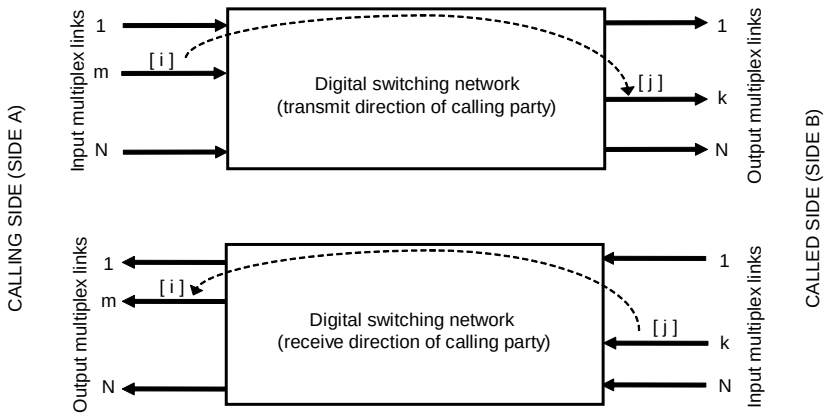
In addition to the line circuits, the Subscriber Access Unit comprises matching units to the digital switching network and the exchange control system, and also to the central testing system controlling the operation of the unit.

b) Switching network

The functional diagram of the digital switching network is shown in Figure 3.2.3. Task of the switching network is to switch through the content of any *input* channel (time slot) to the *output* channel (time slot). It thus follows that the digital switching network has to change positions both in *time* and in *space.* Accordingly, the basic switch modules applied in a digital switching network may be of two kinds:

- *Time switch module* (T-switch) for transposing timing positions, and
- *Space switch module* (S-switch) for transposing space positions.

In the following we consider first the *T-switch*, for interchanging time slots of



Functions of the digital switching network in the transmit direction:
Transfer of the calling voice signal samples from the i-th time slot of the m-th input multiplex link to the j-th time slot of the k-th output multiplex link: m [i] → k [j]

Functions of the digital switching network in the receive direction:
Transfer of the calling voice signal samples from the j-th time slot of the k-th input multiplex link to the i-th time slot of the m-th output multiplex link: k [j] → m [i]

Figure 3.2.3. Draft illustration of digital switching network functions

an input and of an output multiplex path. This means that the *T*-switch is capable of transposiing the content of any (e.g. of the i-th) *input* time slot into any (e.g. in the x-th) *output* time slot. The information arriving in the i-th input time slot has therefore to *wait* in the T-switch as long as the x-th *output* time slot arrives. Note that over the input multiplex path, the information is transmitted periodically in frames so the waiting time within the T-switch cannot be longer than the frame period. Should this not be the case, the waiting information would be overwritten by the information arriving in the next frame, and this would result in a loss of information which is unwanted.

Main functional elements of the *output controlled* T-switch*, as* shown in Figure 3.2.4/a, are the Information Memory IM, the Controlling Memory CM and the Time Slot Counter TSC. The information to be transposed (switched) is sequentially written from the *input* multiplex time slots into the slots of the IM, as assigned by the TSC. Following a time delay defined by the wanted connection, the CM, addressed by the TSC, reads out the IM slots, and transposes their content into the slots of the output

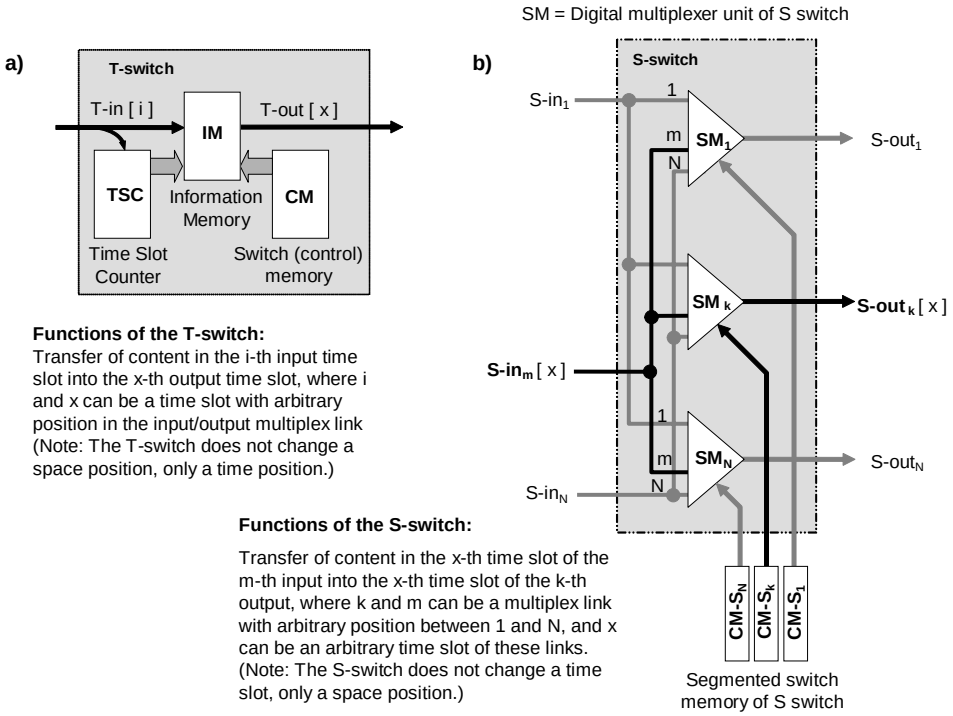


Figure 3.2.4.. Functional layout of digital switching elements

a) T-switch with output control

b) S-switch with output control

multiplex path. This is possible because the CM slot contains the address of an IM slot having a content to be transposed in an *output* time slot, assigned to read out the CM slot. This is possible because the relevant data have to be entered into the CM slots while the connection is established.

The *Space Switch* has the function to change the position of a channel (i.e. the position of the channel time slots) *in space.* It has thus the function to transpose – if required – the content of a time slot in any input multiplex path, to a time slot of the same timing in any output multiplex path, according to an arbitrary combination. Practically this means that a code word arriving in the input x-th time slot will appear again in the x-th time slot at the output, but in another space position. This switchover capability is illustrated in Figure 3.2.4/b showing the architecture and operation of a so-called S-switch *controlled at the output.*

To the input and output of a *symmetrically built* S-switch are connected N digital multiplex paths. These may be either symmetrical or asymmetrical paths. In the example shown in the Figure, both input and output transmission paths are series paths.

Switching is carried out by digital multiplexers denoted by SM. Digital multiplexers have several signal inputs, one address input and a single output. This output is connected to the signal input that is defined by the address input information, as long as the address information is present. According to the Figure, the number of outputs is equal to the number of inputs so the number of digital multiplexers has to be N. Multiplexer inputs of the same serial number are multiplexed.

The multiplexers are addressed, by the segmented CM controlling memory, with the input serial numbers entered into the segment slots, and read out by the TSC. These input serial numbers have to be entered when setting up a connection. Every CM segment is assigned to an SM addressing input, while the slots of the segments are assigned to the time slots of the output multiplex path (similarly to the arrangement used for the CM memory of the T-switch).

Due to restricted space, the switching networks made up of T- and S-switches will not be covered. However, the widely used T-S-T structure will be presented in the followings.

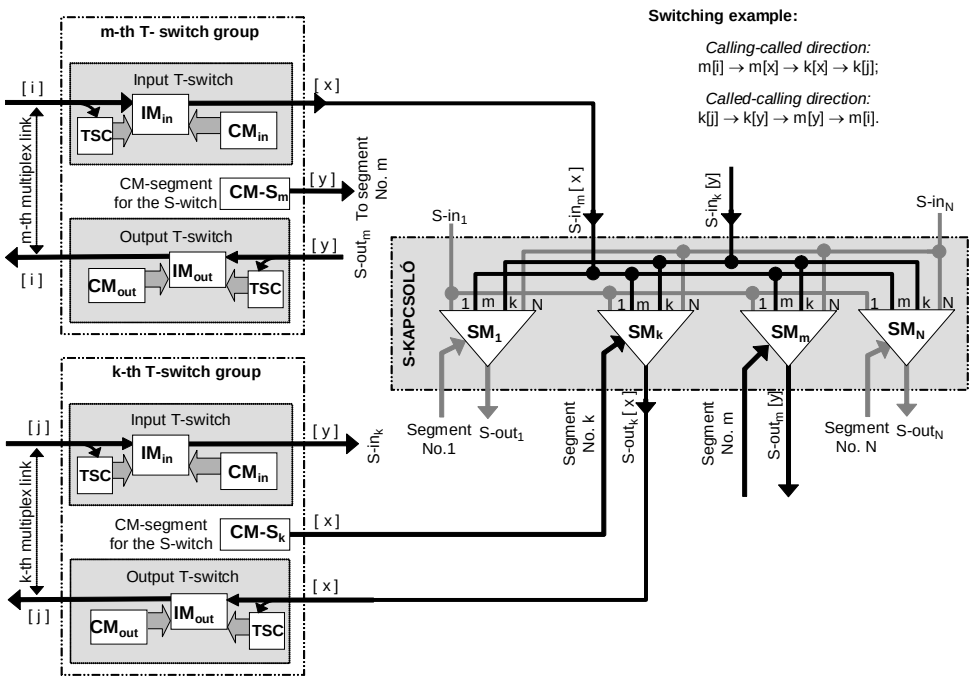


Figure 3.2.5. Functional elements of a digital switching network with T-S-T architecture

We have already outlined the tasks realized by a digital switching network on hand of Figure 3.2.3, and this example will further be followed by explaining the switching network having a T-S-T structure. This is characterized by input and output T-switches, and an S-switch in between, with following tasks:

- *Input T-switches:* to transpose the input multiplex path time slot contents to intermediate internal time slots,
- *S-switch:* to carry out the space switches required by the connection by means of the internal time slots,
- *Output T-switches:* to transpose the internal time slot contents to the selected output time slot.

According to the structure shown in Figure 3.2.5, input/output T-switches belonging to input/output multiplex paths of identical serial numbers, further the control memory of the S-switch are all grouped into a single module. In the Figure, the assumed connection path is separately shown, extending from the input time slot to the output time slot.

3.2.4. Control function

In a digital exchange, the set-up, supervision and disconnection of connections are all under stored program control. Evidently, this task involves the handling of the signalling required for setting up connections. In the followings, the *system structures* of the digital exchanges will be dealt with.

In distributed hierarchically controlled systems the tasks are grouped, and these structured tasks thus defined are distributed between several inter-connected processors.

Several architectures are applied for multi-processor control systems. All structures, proven in practice, have technical and economical advantages and disadvantages. In all cases, the manufacturer is responsible for selecting a structure for his developed system, deciding which advantages should be utilized in his design and which disadvantages should be ignored. This also means that any control structure cannot be unambiguously characterized as being better or worse than others. No such attempts will be made in the following, but rather, the digital exchanges with stored program control in operation in Hungary will be presented, without any evaluation, together with the operating principle of the applied control structures. The following three control structures will be dealt with:

- peripheral bus control,
- control with switched signalling channels,
- distributed control.

a.) Peripheral bus control

The functional diagram of the peripheral bus control principle is shown in Figure 3.2.6. The tasks of the processor systems are allotted according to hierarchical order. a central processor is applied, supported by peripheral processors. The central processor receives the information from these peripheral processors following a “preliminary processing”, and even in the return direction, the central processor has not to process the tasks in every detail as the peripheral processors are helpful in supporting the details of the execution.

One of the main characteristics of this structure is the application of the *peripheral bus.* All instructions, data and other information, required by the operation

of the complete control system, are transmitted only over this bus system. During switch-on or installation of new software versions, the central processor is first loaded with the required programs and database. The central processor thereafter controls the loading of the other processors.

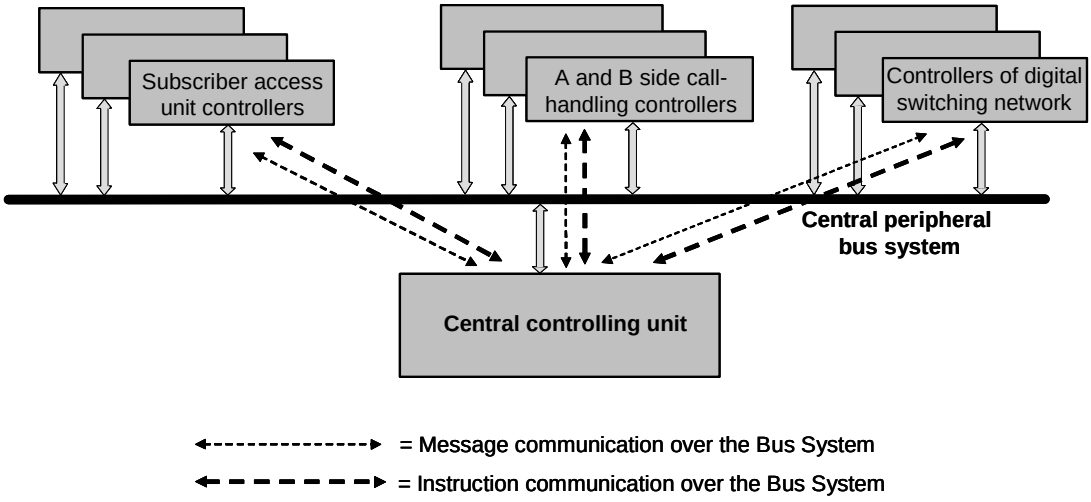


Figure 3.2.6. Functional layout of centralized control with peripheral bus system

b.) Switched signalling channel control

The functional diagram of switched signalling channel control is shown in Figure 3.2.7. This structure is also based on a *hierarchical processor system* but the information exchange between processors is realized via switched signalling channels through a switching network. During switch-on or installation of new software versions, the method of loading the programs and the database is essentially corresponding to the solution given for the peripheral bus system.

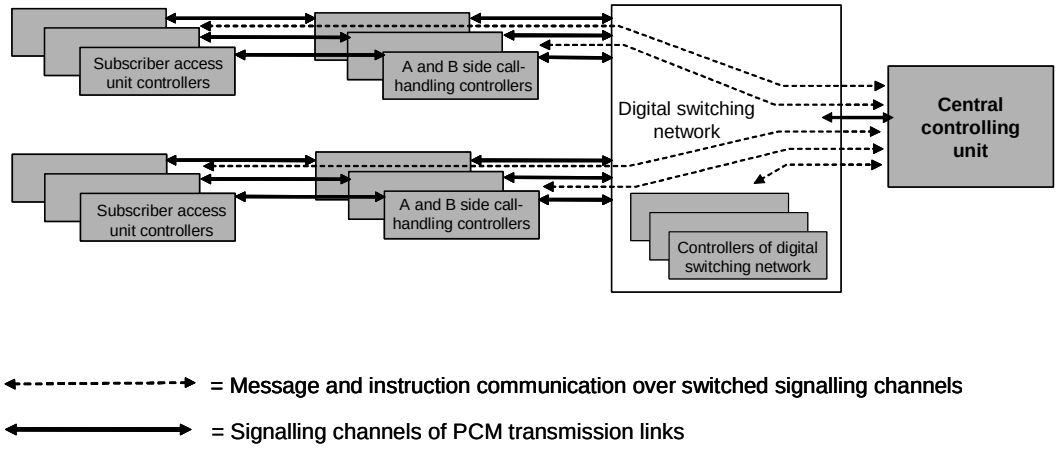


Figure 3.2.7. Functional layout of centralized control with switched signalling channels

c.) Distributed control

The functional diagram of distributed control is shown in Figure 3.2.8. In this structure, there is no central control for supervising all details of the calls (i.e. there is no hierarchy). Every switching unit has its own processor, capable of performing the tasks independently, and handling the user demands while co-operating with neighbouring processors. Call handling is thus provided by co-operation of these processors.

However, even in this kind of control structure, a “common” processor is available but only for co-ordination tasks. This means that only this processor is able to supervise the system administration tasks and the actual status of any switching device. This co-ordination processor is responsible for “organizing” the current processor functions, i.e. to decide which processors will be inter-working to meet the current call requirement. However, this processor is not responsible for actually handling calls or any other user demands. A characteristic feature of the distributed control is *the lack of a peripheral bus,* further the more complicated loading of programs or data during switch-on or installation of a new software version, as compared to the centralized hierarchical structure.

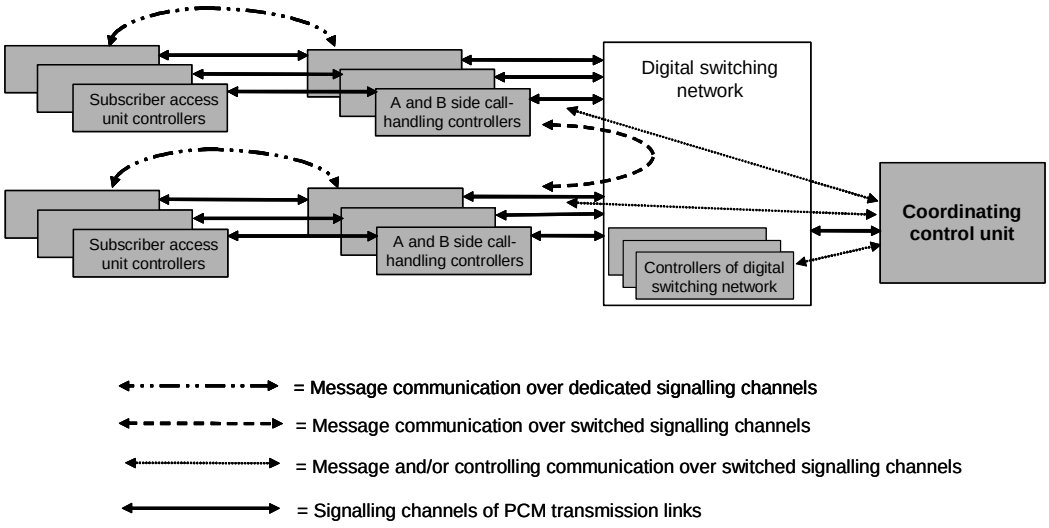


Figure 3.2.8. Functional layout of distributed control

3.2.5. Reliable operation

The electronic components of digital telephone exchanges will fail according to a lower or higher probability, resulting in an unexpected breakdown of the circuits

concerned. The best solution against this malfunction is the instantaneous (or very fast) substitution of these failed circuits within a minimum timespan. Fast substitution is realized by a protection switching system using a hot standby unit. This means that the functional units used for substitution are in an *operational status*. As soon as a failure occurs, an operating unit will substitute the failed unit either automatically or as a result of a manual *remote control command*.

Fault tolerant operation of the units operated under a protection switching system is only possible when only one failure at a time appears in the system. This is why in telephone exchanges, not a single “large” protection is provided for the whole system but the units of the system are individually protected. This means that one unit can tolerate only a single failure but in the complete system, several failures at a time are permitted without a complete breakdown.

In digital exchanges, two methods are used for operating the duplicated units:

- synchronous duplex method
- load sharing method.

According to the *synchronous duplex* method, the system is made up from two functional units operating in parallel, and processing synchronously the input events.

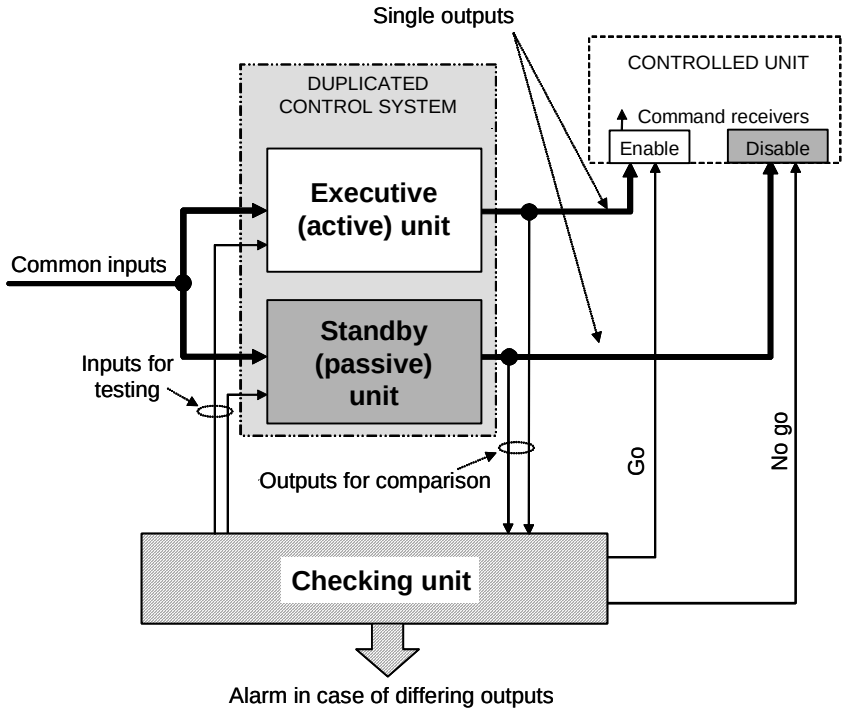


Figure 3.2.9. Principle of reliable operation with duplicated synchronous operation

In the absence of failure, one unit is active, and the other unit is passive. However, only the active unit can execute peripheral control. Figure 3.2.9 shows the functional layout of a system with synchronous duplex units and the controlling environment.

The synchronous duplex system includes a verifying unit that has to notice when the executive and standby units provide different output informations (this would necessarily imply that one of them reveals incorrect operation), and also to automatically ascertain which of the two units failed. Evidently, this cannot perfectly be ascertained but a correct decision with high probability is required (Incorrect decision would imply the breakdown of the duplicated control system). In addition, the verifying unit should also identify the output signals pertaining to the active and passive units, and forward this information to the command receivers as these should only accept information from the executive unit. As soon as a failure is noticed, the verifying unit will separate the failed unit from the control system, activate the good unit, and produce an alarm for the maintenance personnel.

Following separation of the failed unit and declaring the other unit active, the control system should, with high probability, further operate without interruption, in spite of the fact that half of it failed. Evidently, the system will further operate without

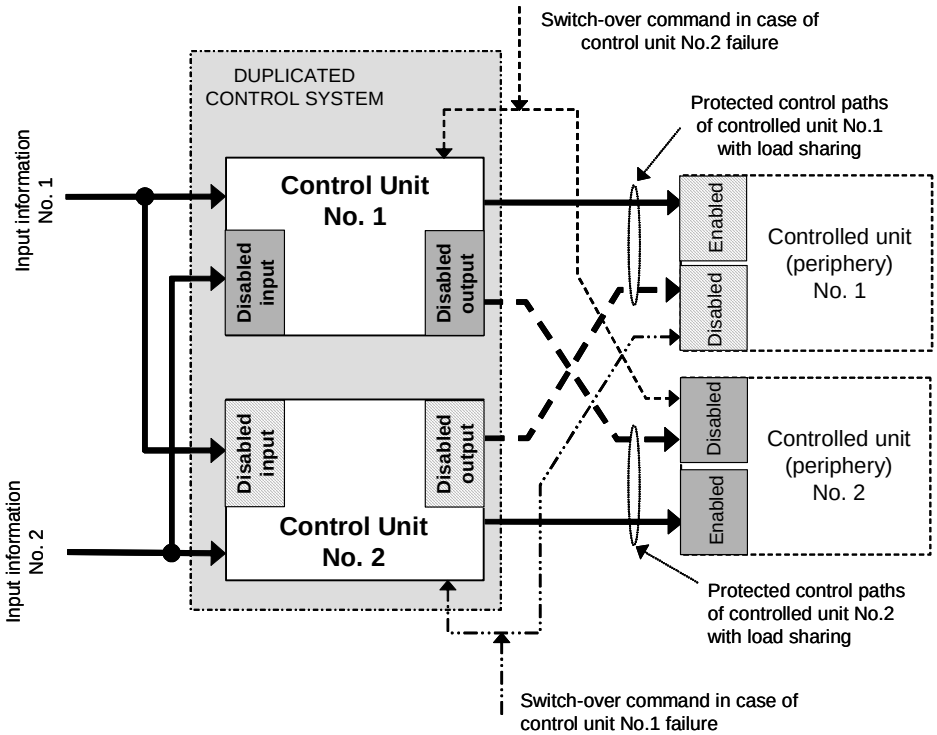


Figure 3.2.10. Principle of reliable operation with load-sharing

standby while the location procedure of the faulty element is in progress. The original condition will be restored only after the faulty element has been replaced and the repaired unit has been put back into operation, following operational tests.

In the case of protection *with load sharing*, at least two, but in some cases more units are operated as mutual standby units. Without failure, all units have separated but similar functions, and each unit operates according to its prescribed task. When one of the units fails then other units will take over the function of the failed unit, and from now on, will perform its previous task with a higher loading.

Figure 3.2.10 shows, in an over-simplified form and confined to the principle only, the functional layout of duplicated control with load-sharing. For simplicity and better understanding, the diagram does not show the environment that is necessary for flawless operation. Evidently, the requirement of having only a single failure at a time is valid here too, so the time span between the occurrence of the failure and the restoration of the original condition should be as short as possible.

Translated by Tamás Sárkány

3.3. Traffic models and teletraffic dimensioning

Sándor Molnár, author

Béla Frajka: reviewer

3.3.1. Introduction

The basic teletraffic principles, equations and an overview of the nature of network traffic are given in Chapter 1.7. Based on these preliminaries this chapter presents the most important traffic models, which are possible candidates to capture the main characteristics of network traffic. Models range from very simple to very complex. In practice, it is a compromise how complex a model we chose to capture more accurately traffic characteristics but also keeping mathematical tractability and computability as simple as possible. With a chosen model applied in a queueing context the aim is generally to find the performance measures in the investigated scenario.

We also provide a survey about the dimensioning principles in both classical telephony and also in data networks with the Internet in the focus.

3.3.2. Traffic models

Traffic consists of single or multiple arrivals of discrete entities (packets, cells, etc.). It can be mathematically described by a *point process*. There are two characterizations of point processes: by the *counting processes* or the *interarrival time processes* [3.3.1]. The counting process $\{N(t)\}_{t=0}^{\infty}$ is a continuous-time, non-negative integer-valued stochastic process, where $N(t) = \max\{n : T_n \le t\}$ is the number of arrivals in the interval (0,*t*]. The interarrival time process is a real-valued random sequence $\{A_n\}_{n=1}^{\infty}$, where $A_n = T_n - T_{n-1}$ is the length of the time interval between the n^{th} arrival from the previous one. The traffic is called compound traffic in case of batch arrivals. In order to characterize compound traffic the batch arrival process $\{B_n\}_{n=1}^{\infty}$ is defined, where B_n is the number of units in the batch. Another useful notion is the

workload process $\{W_n\}_{n=1}^{\infty}$. It is described by the amount of work W_n brought to the system by the n^{th} arrival.

In the following a number of traffic models are described that can be used to generate traffic characterized by sequences of $\{N(t)\}_{t=0}^{\infty}, \{A_n\}_{n=1}^{\infty}, \{B_n\}_{n=1}^{\infty}$ or $\{W_n\}_{n=1}^{\infty}$.

In a **renewal process** $\{A_n\}_{n=1}^{\infty}$ are independent, identically distributed with general distribution [3.3.1], [3.3.2]. This model is simple but non-realistic in many cases because it is not able to capture the strong correlation structure present in most of the actual data traffic.

The **Poisson process** [3.3.1], [3.3.2] is a renewal process whose interarrival times $\{A_n\}_{n=1}^{\infty}$ are exponentially distributed with rate parameter λ. The definition can also be given by the counting process, where $\{N(t)\}_{t=0}^{\infty}$ has independent and stationary increments with Poissonian marginals, i.e. $P\{N(t)=n\} = \exp(-\lambda t)(\lambda t)^n / n!$ Poisson processes are very frequently used in teletraffic theory due to their simplicity and several elegant properties. The voice call arrivals in telephony are typically modeled by Poisson processes.

Bernoulli processes [3.3.1], [3.3.2] are the discrete-time analogs of Poisson processes. In this model the probability of an arrival in any time-slot is *p*, independent of any other one. The number of arrivals in *k* time-slot is binomially distributed, i.e., $P\{N(t)=n\} = \binom{k}{n} p^n (1-p)^{k-n}$ and the times between arrivals are geometrically distributed, i.e., $P\{A_n = j\} = p(1-p)^j$

Phase-type renewal processes [3.3.1], [3.3.2] compose a special class of renewal processes having *phase-type distributed* interarrival times. It is an important class because these models are analytically tractable and, on the other hand, any distribution can be arbitrarily approximated by phase-type distributions.

Markov-based models [3.3.1], [3.3.2] introduce dependence into the random sequence A_n. The construction of the model is the following. Consider a Markov process $M = \{M(t)\}_{t=0}^{\infty}$ with a discrete state space. *M* behaves as follows: it stays in state *i* for an exponentially distributed holding time with parameter λ_i which depends on *i* alone. Then it jumps to state *j* with probability p_{ij}. Each jump of this Markov

process is interpreted as signaling an arrival so the interarrival times are exponential. This is the simplest *Markov traffic model*.

Markov renewal processes [3.3.1], [3.3.2] are more general than simple Markov processes but they can still be handled analytically. A Markov renewal process $R=\{(M_n,\tau_n)\}_{n=0}^{\infty}$ is defined by the Markov chain $\{M_n\}$ and its associated inter-jump times $\{\tau_n\}$, subject to the following constraints: the distribution of the pair (M_{n+1},τ_{n+1}), of next state and inter-jump time, depends only on the current state M_n, but not on previous states nor on previous inter-jump times. In this model arrivals can also be interpreted when jumps occur.

Markov arrival processes (MAP) [3.3.1], [3.3.2], [3.3.3] constitute a broad subclass of Markov renewal processes. In MAP interarrival times are phase-type and arrivals occur at the absorption instants of the auxiliary Markov process. Moreover, the process is restarted with a distribution depending on the transient state from which the absorption had just occurred. MAP is still analytically usable and it is a very versatile process for modeling purposes.

In a **Markov-modulated process** a Markov process is evolving in time and its current state controls the probability law of traffic arrivals [3.3.1], [3.3.2]. Consider a continuous-time Markov process $M=\{M(t)\}_{t=0}^{\infty}$ with state space of *1,2,…m*. While *M* is in state *k*, the probability law of traffic arrivals is completely determined by *k*. When *M* goes to another state, say, state *j*, then a new probability law of traffic arrivals takes effect for the duration of state *j*, and so on. In other words, the probability law of traffic arrivals is modulated by the state of *M*. These stochastic processes sometimes also called *double stochastic processes*. The modulating process can also be much more complicated than a Markov process but such models are less tractable analytically.

The **Markov Modulated Poisson Process** (MMPP) [3.3.1], [3.3.2], [3.3.3] is the most commonly used Markov-modulated traffic model. In this model when the modulating Markov process is in state *k* of *M* then arrivals occur according to a Poisson process at rate λ_k . The simplest case of MMPP is the two-state MMPP model when one state is associated to an “ON” state with a specific Poisson rate, and the other is an “OFF” state with associated rate zero. This model is also called as *interrupted Poisson process*. Such models are used for modeling voice traffic sources

with the ON state corresponds to talk spurt and OFF state corresponds to silence period.

In the **Markovian transition-modulated processes** [3.3.1], [3.3.2] the transition of the Markov process $M = \{M(t)\}_{t=0}^{\infty}$ is the modulating agent rather than the state of *M*. State transitions can be described by a pair of states: the one before transition and the one after it. The number of arrivals B_n in slot *n* is completely determined by the transition of the modulating chain given by $P\{B_n = k \mid M_n = i, M_{n+1} = j\} = t_{ij}(k)$, which is independent of any past state information.

In the **Generally Modulated Deterministic Process** (GMDP) [3.3.3] the source can be any of the possible *N* states. While it is in state *j* the traffic generated at a constant rate λ_j. The time spent in state *j* can be described by a generally distribution but in most of the cases it is assumed geometric in order to keep analytical tractability. If you consider a two state GMDP where one of them has zero generation rate we have the slotted-time version of the ON/OFF model.

In the **fluid traffic modeling** technique the traffic is considered as a fluid instead of individual traffic units [3.3.1], [3.3.2], [3.3.3]. This is a good model where the individual traffic units (e.g. packets) are numerous relative to a chosen time scale. The advantage of this technique is the simplicity compared to traffic models that are aimed to capture the structures of individual traffic units. The simplest types of fluid models are assuming two states: an ON state when traffic arrives deterministically at a constant rate λ, and an OFF state when there is no traffic carried. In order to keep analytical tractability the durations of ON and OFF periods are typically assumed to be exponentially distributed and mutually independent. In other words, they form an alternating renewal process.

Autoregressive traffic models define the next variate in the sequence as an explicit function of the previous variates within a time window stretching from the present to the past [3.3.1], [3.3.2], [3.3.3]. Typical examples of these models are the *linear autoregressive (AR) processes*, the *moving average (MA) processes*, the *autoregressive moving average (ARMA) processes* and the *autoregressive integrated moving average (ARIMA) processes*. These models were found to be useful to characterize VBR video traffic.

The **Transform-Expand-Sample** (TES) [3.3.1], [3.3.2] approach aims to construct a model satisfying three requirements: marginal distributions should match its empirical counterpart, autocorrelation should approximate its empirical counterparts up to a reasonable lag and the sample paths generated by the model should "resemble" the empirical time series. TES models can be used e.g. for constructing MPEG video models.

Fractional Gaussian Noise (FGN) [3.3.1] is an exactly second-order self-similar process with self-similarity parameter *H*, provided *½<H<1*. It is a stationary Gaussian process, $X = \{X_k\}_{k=1}^{\infty}$, with autocorrelation function of the form $\rho_X(k) = \frac{1}{2}(|k+1|^{2H} - 2|k|^{2H} + |k-1|^{2H}), k \geq 1$. FGN is also long-range dependent (LRD) with parameter *H*: $\rho_X(k) \approx H(2H-1)|k|^{2H-2}, k \to \infty$. FGN can be a candidate traffic model for characterizing aggregated LRD traffic at backbone links.

The **fractional ARIMA** (FARIMA) [3.3.1] model is based on the classical ARIMA(*p,q,d*) model but the parameter *d* used with the difference operator is allowed to take fractional values. FARIMA models are more flexible than FGN models to capture LRD traffic because they can also be tuned to capture the short-range dependent (SRD) characteristics as well.

The **M/Pareto model** is a Poisson process with rate λ of Pareto distributed overlapping bursts [3.3.4]. During the burst the arrival process is constant with rate *r*. The burst length period has a Pareto distributions with parameters $1 < \gamma < 2, \delta > 0$: $P\{X > x\} = \begin{cases} \left(\frac{x}{\delta}\right)^{-\gamma}, x \geq \delta \\ 1 \quad , \text{otherwise} \end{cases}$. This model generates LRD traffic with parameter $H = (3-\gamma)/2$, so it is also a good candidate to model fractal traffic.

Based on the various traffic models outlined above one can apply these models or combination of some of them to model specific application traffic. These models are the **applied traffic models to specific applications**. Here we present a guideline for possible modeling alternatives of some popular applications [3.3.5].

Application	Model	Distribution
TELNET	session interarrival times	exponential
	session duration	lognormal
	packet interarrival times	Pareto
	packet size	mostly 1 byte packets
FTP	session interarrival times	exponential
	number of items	empirical
	item size	log-Normal
CBR voice	session interarrival times	exponential
	session duration	exponential
	packet interarrival times	constant
	packet size	constant
VBR video teleconferencing	frame interarrival times	constant
	frame size	Gamma
MPEG video	frame interarrival times	constant
	scene length	Geometric
	frame size	lognormal
WWW	request interarrival times	exponential
	document size	Pareto

Table 3.3.1. Different models describing sevices

3.3.3. Teletraffic dimensioning of classical telephone networks

Teletraffic theory was fundamental to the design of classical telephone networks from the beginning. By assuming a stationary Poisson call arrival process the traffic and performance relationship can be expressed by the well-known Erlang

$$B = E(a,n) = \frac{a^n / n!}{\sum_{i=0}^{n} a^i / i!}$$

loss formula which gives the probability of call blocking *B*, when a certain volume of traffic, *a*, is offered to a given number of circuits, *n*:

It expresses that the blocking probability is a simple measure of the offered traffic. Note that blocking probabilities are insensitive to other details about the nature of traffic such as distribution of call holding time. (The formula is valid for an *M/G/n/n* queueing system.) This famous formula was intensively used during the history of teletraffic theory. Because telephone calls are initiated by independent individuals making independent decisions, random models assumed to be stationary within a busy hour are appropriate for engineering purposes. Since these calls are all point-to-point with a fixed bandwidth the Erlang formula was an excellent guide for network engineering.

Important refinements of this formula were also developed to different networking scenarios but the Erlang loss formula (and also the related Erlang delay formula) are still extensively used by teletraffic engineers in their daily work. No doubt that this formula got the highest success among all the results of teletraffic theory.

Besides the Erlang loss and delay formulae a number of techniques have been developed for telephone networks. These are, for example, the equivalent random method based on the work of Wilkinson, the different descriptions of traffic burstiness by peakedness and indices of dispersion, and models like the Engset model, etc. [3.3.11].

3.3.4. Teletraffic dimensioning of the Internet

We are just in the phase to see the birth of the teletraffic theory of the Internet. Currently network provisioning is based on some rules of thumb and teletraffic theory has no major impact in the design of the Internet. As we discussed in chapter 3 the nature of the data traffic is significantly different from the nature of voice traffic and no general laws can be found as it was in the case of voice traffic. New techniques and models are expected to develop in teletraffic theory of the Internet to cope with these challenges. In the following we review the most possible two alternatives of Internet teletraffic engineering. The one is called “big bandwidth philosophy”, the other is called “managed bandwidth philosophy”.

3.3.4.1 The big bandwidth philosophy

There is a certain belief that there is no real need for some advanced teletraffic engineering in the Internet because the overprovisioning of resources can solve the problems. This is the “*big bandwidth philosophy*”. People from this school say that in spite of the dramatic increase of Internet traffic volume in each year the capacity of links and also the switching and routing devices will be so cheap that overprovisioning of resources will be possible. It is worth investigating a little bit more deeply how realistic the “big bandwidth philosophy” is.

They expect that the transmission and information technology can follow the “Internet traffic doubling each year” trend [3.3.10] and can provide cheap solutions. From a technological point of view it seems that this expectation is not unrealistic at

least in the near future. Indeed, if you imagine today's Internet and you just increase the capacity of links you could have a network which supports even real-time communications without any QoS architectures. The current best effort type Internet could do it!

On the other hand, the locality of data in the future will also be dominant, which makes caching an important technical issue of future networks [3.3.10]. Even today if you want to transmit all the bits that are stored on hard drives it would take over 20 years. This trend probably gives a relative decrease in the total volume of transmitted information.

Another important factor is that the streaming traffic, which really requires some QoS support is not dominant in the Internet [3.3.10]. It was believed that it would become dominant but none of these expectations have been fulfilled so far, which can also be predicted for the future. The demand of this traffic type is not growing as fast as the capacity is increasing. Consider the following example: we have 1% streaming traffic so it needs some QoS support. We have two options. We can introduce some QoS architecture or we can increase the capacity by 5%. The people from the "big bandwidth philosophy" school argue that the second one is cheaper. They also say that multimedia applications will use *store-and-reply* technique instead of real-time streaming. They argue that the capacity of magnetic storage is increasing at about the same rate as transmission capacity. Moreover, due to transmission bottlenecks (e.g. wireless link) it makes sense to store information in local.

It is also interesting if we investigate the reason of capacity increase in the previous years. For example, we can see that people are not paying for cable modems or ADSL because their modem links could not bring them more data, but because when they click on a hyperlink they want that page on their screen immediately! So they need the big capacity not for downloading lots of bits but rather achieving a *low latency* when they initiate a file download. This is also the reason for the fact that the average utilization of LANs have been decreased by about a factor of 10 over the last decade: people wanted high bandwidth to provide low latency!

Will overprovisioning be the solution? Nobody knows at this time. It is rather difficult to predict what will happen mainly because this is not only a technical issue but rather depends on political and economic factors. However, as a modest

prediction we might say that even if overprovisioning can be a solution for backbone networks it is less likely that it will happen also in access networks. For cases where overprovisioning cannot be applied we have a limited capacity which should be managed somehow. This leads us to the second alternative which is the *"managed bandwidth philosophy"*.

3.3.4.2 The managed bandwidth philosophy

In the case of limited network resources some kind of traffic control should be implemented to provide appropriate capacity and router memory for each traffic class or stream to fulfill its QoS requirements. Basically, there are three major groups of QoS requirements: transparency, accessibility and throughput [3.3.7]. *Transparency* expresses the time and semantic integrity of transferred data. As an example for data transfer semantic integrity is usually required but delay is not so important. *Accessibility* measures the probability of admission refusal and also the delay for set up in case of blocking. As an example the blocking probability is in this class, which is a well-known and frequently used measure in telephone networks. The *throughput* is the main QoS measure in data networks. As an example in today Internet a throughput of 100Kbit/s can ensure the transfer of most of the web pages quasi-instantaneously (less than one second).

Considering the traffic types by nature two main groups can be identified: stream traffic and elastic traffic [3.3.7]. The *stream traffic* is composed of flows characterized by their intrinsic duration and rate. Typical examples of stream traffic are the audio and video real-time applications: telephone, interactive video services, and videoconferencing. The time integrity of stream traffic must be preserved. The negligible loss, delay and jitter are the generally required QoS measures.

The *elastic traffic* usually consists of digital objects (documents) transferred from one place to another. The traffic is elastic because the flow rate can vary due to external causes (e.g. free capacity). Typical elastic applications are the web, e-mail or file transfers. In case of elastic traffic the semantic integrity must be preserved. Elastic traffic can be characterized by the arrival process of requests and the distribution of object sizes. The throughput and the response time are the typical QoS measures in this class.

In the following two subsections we overview the main principles of managing stream and elastic traffic, respectively.

3.3.4.3 The open-loop control of stream traffic

The stream traffic is usually controlled by an *open-loop preventive traffic control* based on the notion of traffic contract [3.3.7]. Traffic contract is a successful negotiation between the user and the network in which user requests are described by a set of traffic parameters and required QoS parameters. Based on these requests the network performs an admission control accepting the communication and the traffic contract only if QoS requirements can be satisfied.

The effectiveness of this control highly depends on how accurately the performance can be predicted based on the *traffic descriptors* [3.3.6]. From the practice it turned out that it is not simple to define practically useful traffic descriptors. It is because it should be *simple* (understandable by the user), *useful* (for resource allocation) and *controllable* (verifiable by the network). Results of intensive research on finding such traffic descriptors with all these properties showed that it is practically impossible. As an example the standardized *token bucket type descriptors* (both in ATM and Internet research bodies) are good controllable descriptors but they are less useful for resource allocation. The users are encouraged to use mechanisms (e.g. traffic shaping) to ensure declared traffic descriptors. Mechanisms can also be implemented at the network ingress to police traffic descriptors (traffic policing). Both shaping and policing are frequently based on the mentioned token bucket type mechanisms.

The major types of open-loop traffic control (admission control) strategies depend on whether statistical multiplexing gain is aimed to be utilized and to what extent [3.3.7]. The following table shows the main categories:

Approach	Buffer sharing	Bandwidth sharing
peak rate allocation	NO	NO
Rate envelope multiplexing	NO	YES
Rate sharing	YES	YES

If no multiplexing gain is targeted to achieve we have the simplest case and we can simply allocate the maximal rate (peak rate) of all the connections, which is called *peak rate allocation*. The advantage of this approach is that the only traffic

descriptor is the peak rate of the connection. The admission control is very simple: it only has to check that the sum of required peak rates is over the total capacity or not. The main disadvantage of peak rate allocation is the waste of resources because statistically it is only a small fraction of the time when all the connections actually transmit traffic at the peak rate.

If we design to share the bandwidth but not to share the buffer among connections, we have the *rate envelope multiplexing* case [3.3.6], [3.3.7], [3.3.8]. This approach also called *bufferless multiplexing* because in the fluid modeling framework of this method no need for a buffer. Indeed, in rate envelope multiplexing the target is that the total input rate is maintained below the capacity. The events of exceeding the capacity should be preserved below a certain probability. i.e., $P(\Lambda_t > c) < \varepsilon$, where Λ_t is the input rate process, *c* is the link capacity and ε is the allowed probability of exceeding the capacity. In actual realizations buffers always needed to store packets which arrive simultaneously (cell scale congestion). All the excess traffic is lost, the overall loss rate is $E[(\Lambda_t - c)^+ / E(\Lambda_t)]$. The loss rate only depends on the stationary distribution of Λ_t and not on its time dependent properties. It is important because it means that the correlation structure has no effect on the loss rate. Therefore the very difficult task of capturing traffic correlations (e.g. long-range dependence) is not needed. The traffic structure can have impact on other performance measures but these can be neglected if the loss rate is small enough. For example, LRD traffic can yield to longer duration of the overloads than SRD traffic but using a small loss rate it can be neglected in practice. The main disadvantage of rate envelope multiplexing is that the utilization is still not very good.

If we want to further increase the link utilization we have to share the buffer as well, see Figure 3.3.1. This is the *rate sharing* method [3.3.6], [3.3.7], [3.3.8] or also called *buffered multiplexing*. The idea here is that by providing a buffer we can absorb some input rate excess. The excess of the queue length in the buffer at some level should be preserved below a certain probability, i.e., $P(Q > q) < \varepsilon$, where *q* is the targeted queue length level, *Q* is the actual queue length and ε is the allowed probability level of exceeding the targeted queue length. In this method much higher multiplexing gain and utilization can be achieved.

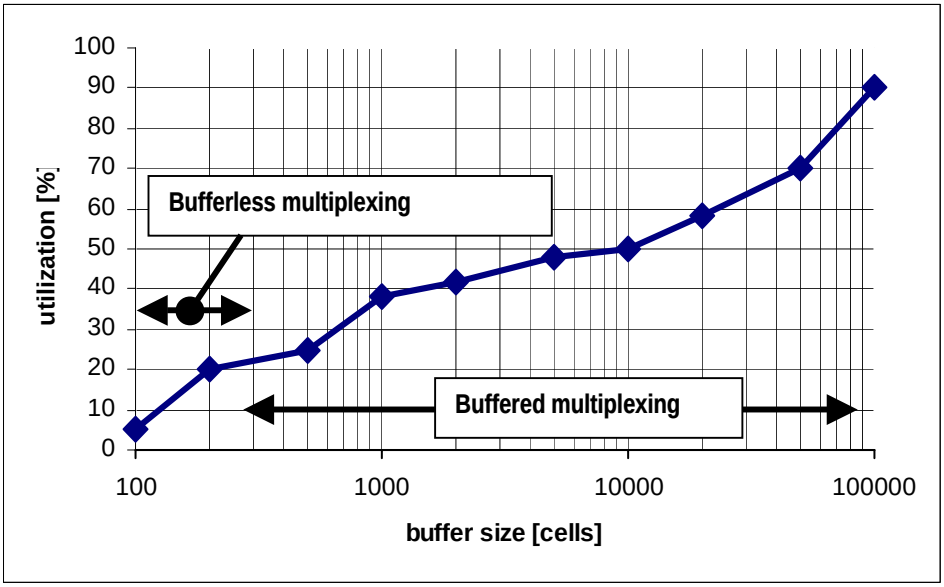


Figure 3.3.1: Alternatives for statistical multiplexing

The main problem in rate sharing is that the loss rate realized with a given buffer size and link capacity depends in a complicated way on the traffic characteristics including also the correlation structure. As an example the loss and delay characteristics are rather difficult to compute if the input traffic is LRD. This is the reason that the admission control methods are much more complicated for rate sharing than for rate envelope multiplexing [3.3.8]. Moreover, the disadvantage is not only the complex traffic control but the achievable utilization is also smaller in case of

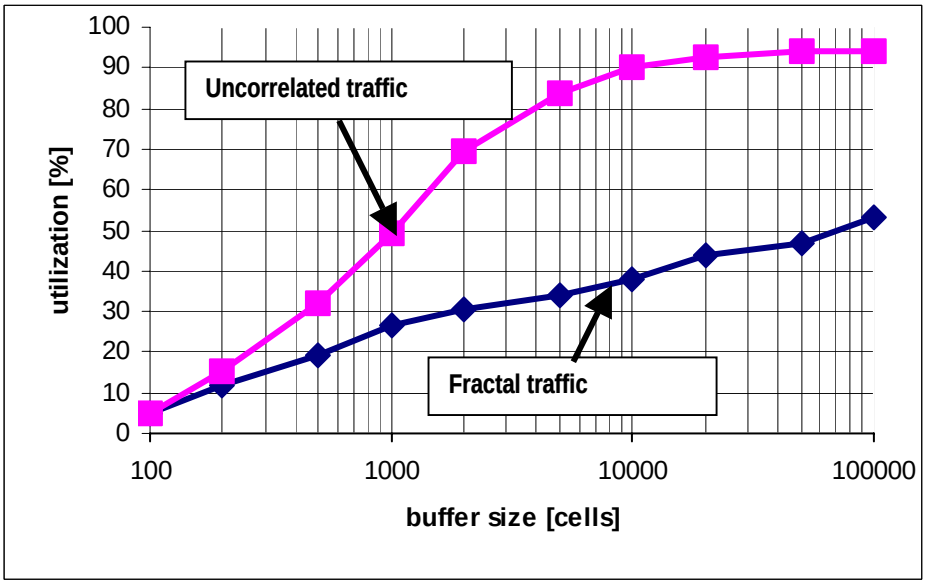


Figure 3.3.2: The impact of correlation structure

fractal traffic with strong SRD and LRD properties, see Figure 3.3.2.

A large number of admission control strategies have been developed for both rate envelope multiplexing and rate sharing [3.3.8]. It seems that the most powerful scheme is a kind of measurement-based admission control where the only traffic descriptor is the peak rate and the available rate is estimated in real-time.

3.3.4.4 The closed-loop control of elastic traffic

Elastic traffic is generally controlled by *reactive closed-loop traffic control* methods [3.3.6], [3.3.7]. This is the principle of the TCP in the Internet and the ABR in the ATM. These protocols target to fully exploit the available network bandwidth while keeping fair shares between contending traffic flows. Now we investigate the TCP as the general transfer protocol of the Internet. In TCP an additive increase, multiplicative decrease congestion avoidance algorithm has been implemented. If there is no packet loss the rate increases linearly but the packet transmission rate is halved whenever packet loss occurs. The algorithm tries to adjust its average rate to a value depending on the capacity and the current set of competing traffic flows on the links of its paths. The available bandwidth is shared in a roughly fair manner among the TCP flows.

A simple model of TCP [3.3.9], which also captures the fundamental behavior of the algorithm, is the well-known relationship between the flow throughput *B* and the packet loss rate *p*:

$$B(p) = \frac{c}{RTT\sqrt{p}},$$

where *RTT* is the TCP flow round-trip time and *c* is a constant. It should be noted that this simple formula is valid in case of a number of assumptions: *RTT* is constant, *p* is small (less than 1%) and the TCP source is greedy. The TCP mechanism is also assumed to be governed by the fast retransmit and recovery (no timeouts) and the slow-start phase is not modeled. More refined models were also developed but the square-root relationship between *B* and *p* seems to be a quite general rule of TCP.

Implementing admission control schemes for elastic traffic is a current and open research issue [3.3.6], [3.3.7]. In such a method the admittance threshold

should be small enough to avoid flow rejection in normal load situations but large enough to ensure satisfactory throughput for admitted flows in overload.

3.3.5. Concluding remarks on teletraffic dimensioning

The importance of choosing a good traffic model determines how successful we are in capturing the most important traffic characteristics. The traffic model applied in the investigated teletraffic system, which is in most of the cases a queueing system, is the complex teletraffic model under investigation. The basic question is the fundamental relationship between the traffic characteristics, network resources and performance measures. Queueing models with some types of traffic models (e.g., Poisson, MMPP, MAP, etc.) are analytically tractable but others (e.g., ARIMA, TES, FGN, etc.) are not. It is a current research issue to develop new theoretical and applied tools to assist in solving teletraffic systems with emerging new and complex traffic models.

Our survey about the dimensioning methods shows that the teletraffic dimensioning of the Internet is not a fully solved problem and several issues are in the scope of current teletraffic research. As opposed to the dimensioning of telephone networks, which can be considered as a well understood and solved issue, the teletraffic theory of the Internet with dimensioning methods is mainly the topic of the future.

References

[3.3.1] D. L. Jagerman, B. Melamed, W. Willinger: Stochastic modeling of traffic processes, In J. Dshalalow, ed., Frontiers in Queueing: Models, Methods and Problems. CRC Press, 1997. pp. 271-320.

[3.3.2] V. S. Frost, B. Melamed: Traffic Models for Telecommunications Networks, IEEE Communications Magazine, March 1994. pp 70-81.

[3.3.3] G. D. Stamoulis, M. E. Anagnostou, A. D. Georganas: traffic sources models for ATM networks: a survey, Computer Communications, vol. 17, no. 6, June, 1994. pp. 428-438.

[3.3.4] R. G. Addie, M. Zukerman, T. D. Neame: Broadband Traffic Modeling: Simple Solutions to Hard Problems, IEEE Communications Magazine, August 1998. pp. 88-95.

[3.3.5] B. O. Lee, V. S. Frost, R. Jonkman: NetSpec 3.0 source Models for telnet, ftp, voice, video and WWW traffic, 1997.

[3.3.6] J. Roberts, Traffic Theory and the Internet, IEEE Communications Magazine, January 2000.

[3.3.7] J. Roberts, Engineering for Quality of Service, in the book of Self-Similar Network Traffic and Performance Evaluation, (eds. K. Park, W. Willinger), Wiley, 2000.

[3.3.8] J. Roberts, U. Mocci, J. Virtamo (eds.), Broadband Network teletraffic, Springer-Verlag, 1996.

[3.3.9] J. Padhye et al. Modeling TCP Throughput: A Simple Model and Its Empirical validation, Proc. SIGCOMM'88, ACM, 1998.

[3.3.10] A. Odlyzko: The history of communications and its applications for the Internet, available at http://www.research.att.com/~amo/doc/complete.html, 2000.

[3.3.11] H. Akimaru. K. Kawashima: Teletraffic, Theory and Applications, Springer-Verlag, 1999.

3.4. Common channel signalling concept

Gusztáv Adamis, author

Gyula Csopaki, reviewer

The conventional channel-associated signalling systems are not suitable for the ISDN. The transmission of speech (data) and signalling requires different features from the transmission medium. In speech transmission, for a relatively long time, we have to transmit a lot of information in a strictly scheduled way (1-1 octet long samples in every 125 us). The signals do not require so strict scheduling, their information content is relatively small, and typically there is quite a lot of time between them while the channel is idle. This gave the idea to separate the voice/data network from the signalling. On the signalling network while we are waiting for the preparation of the next signal of a conversation we can send signals of other conversations on the same signalling link. If we use digital format, then the signals may have a lot of and flexible parameters, and this way the same signalling system can fulfil the requirements of different services.

ISDN and signalling networks must meet different reliability requirements. Because while one failure in the ISDN network can hurt only one or several conversations, but a lost signalling message may e.g. cause that an ISDN connection would "never" released, making much larger loss. By separating the ISDN and the signalling networks we can provide the required reliability of the signalling network without providing unnecessarily the same for the ISDN network. The separated signalling network can also be suitable for transmission of some network management and maintenance information.

But we have to mention some disadvantages, too. The maintenance of a separate network requires extra cost. The structure of the signals will be more complicated, since they have to explicitly identify the voice connection they refer to, and it will be no longer sure that the voice path that was established by the signalling network is really continuous, it should be checked. (The latter two were no problems in the channel-associated signalling, since if the path establishing signals arrived, the

path was continuous, and these signals referred to the conversation would take place on the same path where the signals came.) But the advantages are greater, so the common channel signalling concept has increasing importance.

Now let us see some terminology! On Figure 3.4.1 the common channel signalling concept can be seen. Each switch contains a functionally separated unit, that handles the signalling system. They are the Signalling Points (marked by circles). As can be seen on the figure, the signalling connections may be different to the voice paths, and some extra elements, called Signalling Transfer Points (marked by squares) may also be included.

The CCITT/ITU-T has standardised a common channel signalling system, called CCSS7 (Common Channel Signalling System No. 7). The simplified architectural overview of the CCSS7 can be seen on Figure 3.4.2. Its structure is also hierarchical, but the hierarchical layers are not identical to those in the OSI model, so they are called as “level”s and not as “layer”s.

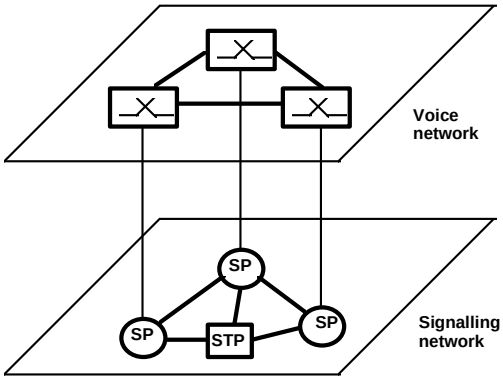


Figure 3.4.1. Common channel signalling concept

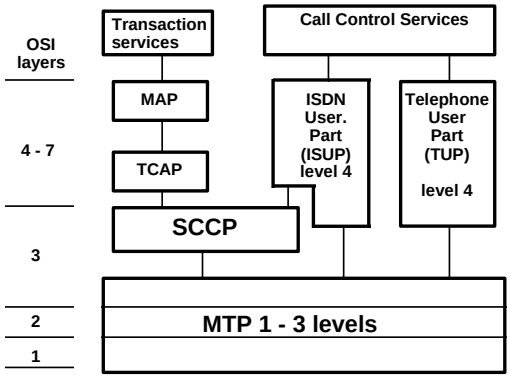


Figure 3.4.2. CCSS7 architecture

The lowest three levels, the Message Transfer Part (MTP) is responsible for the transmission of the signalling information between any signalling point of a signalling network. On the fourth level there are the User Parts (UPs), that generate the signalling messages for the transaction services (e.g. GSM) and for the call control services (e.g. ISDN). The purpose of the SCCP is to maintain logical signalling connections especially for transaction services.

As it was mentioned the signalling network must be highly reliable. This is why it is necessary to have more than one signalling paths between any two signalling

points. Figure 3.4.3. shows the network topology must be established between any two (A and F on the Figure) signalling points in the international network. The essence of the network is that there must be - at least - two possible continuing route from any signalling point, (e.g. AB and AC from A, BD and BE from B, and so on) and - under normal circumstances - 50-50% load sharing must be applied on these routes, that is one half of the traffic between A and F must be routed on AB, and the other half to AC. For increasing the reliability there are transversal connections between STPs (BC and DE). On these routes - under normal circumstances - there is no traffic in the A-F direction, but it does not mean that they are idle, since these links may be parts of other similar networks, too.

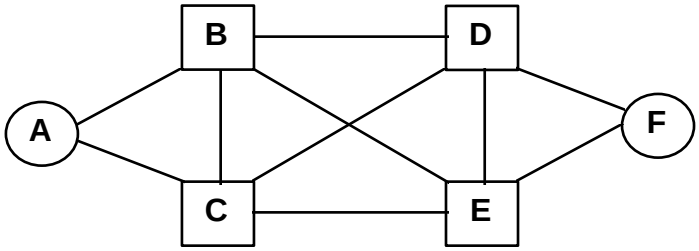


Figure 3.4.3. Recommended signalling network

MTP-1: Signalling Data Link level

defines the physical, electronical features of the signalling network, the shapes of the signals, the access methods etc. The CCSS7 is optimised for the unrestricted, transparent, full-duplex, 64 kbps digital channels, but it is possible to use other - even analogous - media.

MTP-2: Signalling Link level

is designed to provide the errorfree transmission of signalling information between neighbouring signalling points. For this purpose it adds the following signalling link control to the information arrives from the higher level (Figure 3.4.4):

- F: Flag, value: 01111110. Purpose: the separation of signalling messages. To avoid the appearance of the same bit pattern inside the messages the well known bit stuffing technique is used: a 0 is inserted after every five consequent 1s.
- BSN: Backward Sequence Number. The sequence number of the lastly received signalling message (modulo 128). It is a positive acknowledgement for that and all the previous messages.

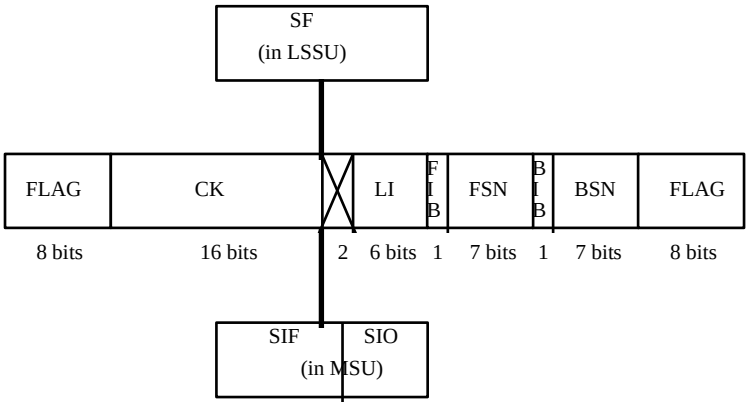


Figure 3.4.4. MTP-2 message structure

- BIB: Backward Indicator Bit. If a not valid message is received, the invertation of this bit can indicate a negative acknowledgement.
- FSN: Forward Sequence Number of the sent message (modulo 128).
- FIB: Forward Indicator Bit. By inverting this bit a repeated message can be indicated.
- LI: Length Indicator. It counts the length of the information field in octets. The signalling message types can be distinguished by the value of this field.

The FISU (Fill-In Signal Unit) does not contain an information field. It must be sent if there is no "useful" message to be transmitted.

In the periodically sent 1 (or 2) octet long LSSU (Link Status Signal Unit) the information field is called as Link Status Field, and indicates the state of the given link (normal, out-of-alignment, congested, etc.)

The third message type is the MSU (Message Signal Unit), that contains "useful" information from the higher levels. In this case the information field is divided into two parts, the Service Information Octet (SIO) chooses the User Part to the message is sent, and the Signalling Information Field (SIF) contains the Routing Label (address) and the message itself.

- CK: 16 bit checksum.

As it can be seen MTP-2 uses the well known sliding window mechanism with explicit negative acknowledgement feature.

MTP-3: Signalling Network level

has two main functions: Signalling Message Handling (SMH) and Signalling Network Management (SMN). SMH is responsible for transmitting the signalling

messages between any two signalling points within the same network ("addressing") while SMN is responsible for the reconfiguration of the signalling network after a failure, insertion of new elements, etc.

Signalling Message Handling

Thefunctional overview of the SMH can be seen on Figure 3.4.5. The message discrimination function decides if an incoming message was sent to the given signalling point, or must be transmitted further. If the incoming message is for the given SP, then the "destination" user part is chosen by the signalling message distribution function. The message routing function is responsible for transmitting the messages sent by a user part of the given SP, and for those that was decided by message discrimination to transmit further.

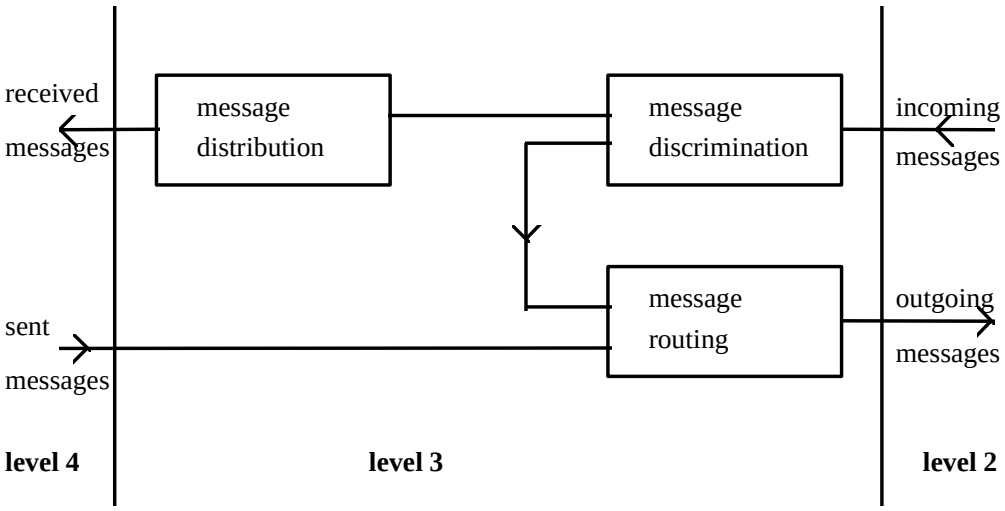


Figure 3.4.5. MTP-3 Message handling

For routing the signalling messages the following addressing scheme is used (Figure 3.4.6). The mentioned service information octet (SIO), that chooses the "destination" user part is followed by the so called Routing Label (RL). The Routing Label is divided into three parts: the Destination Point Code (DPC), the Originating Point Code (OPC) are 14-14 bits long, while the Signalling Link Selection code (SLS) consists of 4 bits. The latter is used for choosing one of the maximum 16 different routes (that are used for increasing the reliability of the network) that may exist between two SPs.

Signalling Link Selection (SLS)	Destination Point Code (DPC)	Originating Point Code (OPC)	Signalling Information Octet (SIO)
4 bits	14 bits	14 bits	8 bits

Figure 3.4.6. MTP-3 Routing label

Signalling Network Management

The SNM has three major parts (Figure 3.4.7). The task of the Signalling Link Management (SLM) is to detect the changes in the availability of the links by the help of the bit error rate measuring function of MTP-2, or by the commands of the Management (operators), and to report these changes to the Signalling Traffic Management (STM). The task of the STM is to keep track on the states of the signalling links and routes originating from the given signalling point, and by the help of the information obtained from the Signalling Link Management to modify its routing tables. The third functional unit is the Signalling Route Management (SRM). It is responsible for distributing the changes made by the STM to those Signalling Points that are not involved directly in the change.

Let us see the structure of the Signalling Traffic Management in details!

The Link Availability Control function stores the states of the links in tables. This is the process that modifies the entries of these tables when necessary according to the information received from Signalling Link Management and initiates the necessary process for modifying the routing of the Signalling Point. The function of the Route Availability Control is similar, but for the routes, not for the links. The role of the Signalling Route Control is, if a change on the routing of the Signalling Point involves or may involve other Signalling Points - apart from those that are directly involved - it initiates the necessary functions (Traffic Prohibited and Traffic Allowed) of the Signalling Route Management. The task of the Signalling Traffic Reconfiguration and Flow Control is to actually initiate the route modification functions, that are the followings:

- Changeover: to divert the whole traffic from a not available link to an available one.
- Changeback: to redirect the traffic to a link became available.

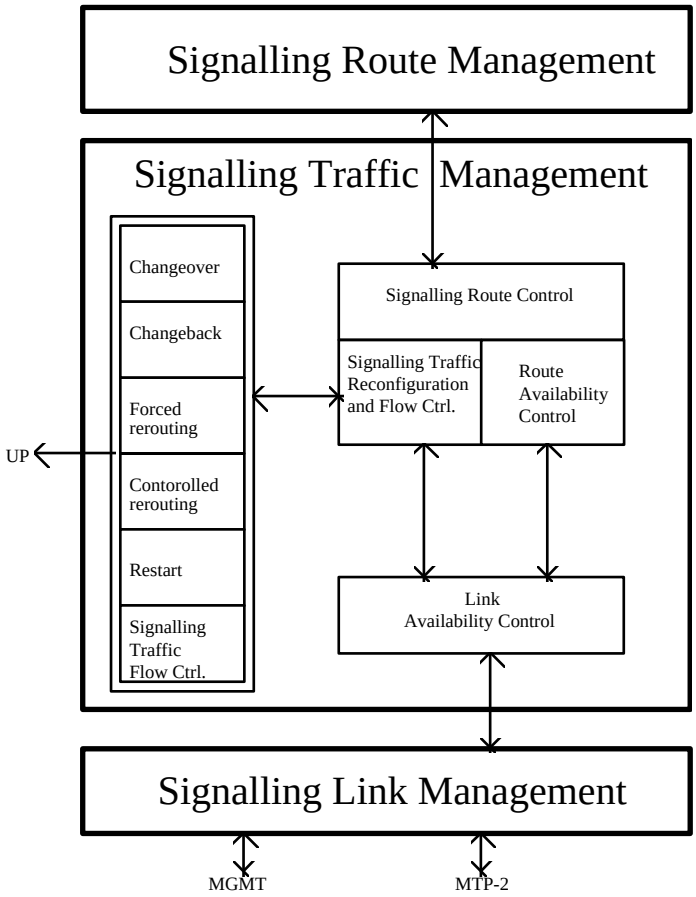


Figure 3.4.7. MTP-3 Signalling Network Management

- Forced Rerouting: to divert the traffic from a not available route to an available one.
- Controlled Rerouting: to redirect the traffic to a route became available.
- Restart
- Signalling Traffic Flow Control: to ask the User Parts to reduce their traffic when congestion.

Level 4: The ISDN User Part (ISUP)

Till this point we discussed how CCSS7 can transmit the signalling information from an SP to an other. Now let us see a User Part that generates and receives these messages. Let us choose the ISUP, that serves the signalling needs of the ISDN/telephony network (Figure 3.4.8.).

Apart from the Call Processing we need a (voice/ISDN) Circuit Supervision Control to be able to manage the voice circuits via the signalling network. The purpose of the other two blocks is obvious.

Structure of the ISUP signals

The functionality of most of the ISUP messages is similar to those of the channel associated signalling. The most important differences are the digital format and the wide parameter set. An illustration of the general format of an ISUP message can be seen on Figure 3.4.9.

Each ISUP message contains the already mentioned routing label and a circuit identification code. The latter identifies the voice/ISDN circuit the signal refers to. It is followed by the message type code. Each message may contain several parameters. Some of them are mandatory (e.g. the dialled number in an Initial Address Message [IAM]), the rest is optional. The length of a parameter may be fixed or may be variable, that is its length may be different in different messages (a typical example is the dialled number). So we can have mandatory fixed and mandatory variable parameters.

The mandatory parameters may be followed by optional ones. The presence of these parameters depends on the situation, e.g. if a value added service is involved in a call. An example for an optional parameter is the number of the calling party in IAM. The whole message is terminated by an End of Optional Parameters field.

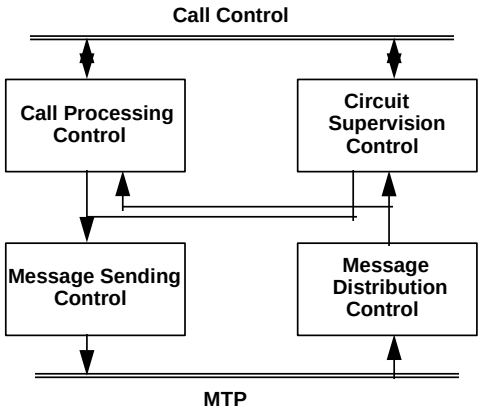


Figure 3.4.8. Functional Blocks of ISUP

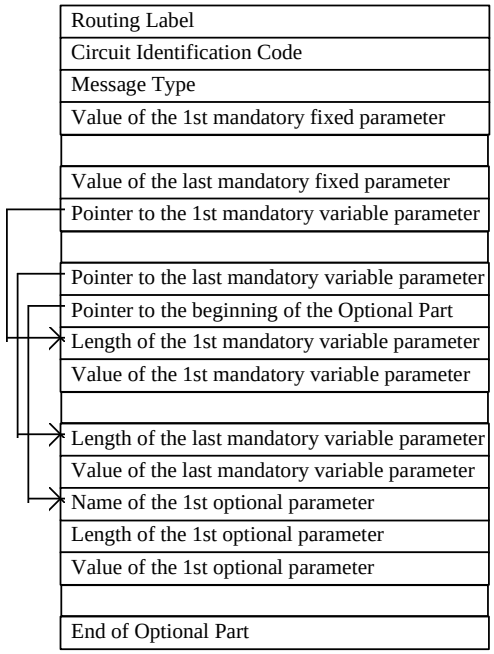


Figure 3.4.9. ISUP message format

Classification of ISUP signals

The ISUP messages can be divided into four functional groups:

1.Call establishment, supervision and release messages

2. Call modification messages

3. Circuit and circuit group supervision messages

4. End-to-end messages (not used in practice)

Most important call establishment, supervision and release messages

- IAM - Initial Address Message - the first message of a connection that implicitly reserves the voice circuit (forward)
- SAM: - Subsequent Address Message - if the dialled number is transmitted by overlap method, it carries the second etc. part(s) of the number (forward)
- ACM - Address Complete Message (backward)
- CPG - Call Progress - used e.g. in call forwarding (backward)
- ANM - Answer Message - sent when called party answered, when receiving, the charging of the call starts (backward)
- REL - Release (any direction)
- RLC - Release Complete (any direction)

Most important call modification messages

- FAR - Feature Activation Request
- FAA - Feature Activation Accepted
- FRJ - Feature Activation Rejected

Most important circuit supervision messages

- CCR - Call Continuity Request (for the voice path)
- LPA - Loop Acknowledge (acknowledge to CCR)
- OLM - Overload Message
- SUS/RES - Suspend/Resume
- BLO/BLA - Blocking request/Acknowledge
- UBL/UBA - Unblocking request/Acknowledge

References:

[3.4.1] CCITT Specifications of Signalling System No. 7.
Recommendations Q.704. - Q.705.

[3.4.2] ETSI Integrated Services Digital Network: Application of the ISDN user part of CCITT Signalling System No. 7. for international ISDN interconnections - CCITT Recommendation Q.767.

3.5. TCP/IP Networks and IP Telephony

Róbert Szabó, author

György Réthy, reviewer

3.5.1. Definition

"Internet telephony (IPT) is transport of telephone calls over the Internet, no matter whether traditional telephony devices, multimedia PCs or dedicated terminals take part in the calls and no matter whether the calls are entirely or only partially transmitted over the Internet." by Jiri Kuthan.

3.5.2. Overview

The main driving force behind IPT is *money-saving* and *easy implementation of innovative services*. Money saving can be reasoned from the point of using a single infrastructure for providing both, Internet access and Internet telephony. Here, however two traditionally competing networking approaches (packet and circuit switched models) are merging toward a single networking infrastructure bearing certain attributes of both ancestors. The biggest packet switched network, i.e., the Internet, is even now facing a problem called quality of service (QoS) assurance (See section 3.7) that has long been solved in the Plain Old Telephone Systems (POTS). On the other hand only packet switches could effectively support data transport as well as packetized voice. As for money saving, packet networks benefit of effective multiplexing which results in better utilization of transmission capacities unlike the over-provisioned POTS networks where circuits are either used for voice or for nothing. In POTS to maintain low call blocking networks are generally over dimensioned. Furthermore, not only the providers, but also their clients will profit of lower costs eventually.

Another aspect of money saving comes from the flat rating of Internet versus the hierarchical rating of Public Switched Telephone Network (PSTN). In nowadays Internet communication distances do not matter but only the data amount, connection

time or pre-negotiated charges. This way letting long-distance calls be routed over the Internet significantly reduces costs. Nevertheless, future predictions say that IPT and PSTN prices will equalize as the provided quality of service converges [3.5.5], which naturally means improvements of the Internet.

Ease of implementation can be argued by the software-oriented nature of IPT evolution. Software-oriented solutions have the benefit to easily extend and integrate with other services and application, e.g. white boarding, voice mails, WWW based call centers, etc…

In spite of the promising benefits of IPT, its wide development is still hindered by the inconvenient QoS support in the bearer transport, i.e., the Internet. Additional technical issues like accounting, billing, security, charging, roaming are still open yet. Further, dial-in access does not work well with IPT; however always-connected services like CableTV, xDSL lines etc… may promote it.

3.5.3. Standards and Architectures

Nowadays the dominant specification for IPT is the ITU's[1] H.323 recommendation, which includes a series of recommendations to describe system components, control messages, procedures among components and services for multimedia communications over IP-based networks. On the other hand, the IETF[2] is also developing a protocol named Session Initiation Protocol (SIP) for multimedia session initiation. In the SIP protocol family the technical details are carried in the session descriptor protocol (SDP) [3.5.9]. SIP is used to initiate sessions among users by providing user location services, call establishment and call participant management. Both protocols utilize the Real-Time Transport Protocol (RTP)[3.5.6]. H.323 terminals use RTP as the transport protocol for multimedia, and SIP was designed to control multimedia sessions delivered over RTP. Nevertheless, other protocols can be as well utilized like TCP or pure UDP. [3.5.1][3.5.2]

[1] International Telecommunication Union

[2] Internet Engineering Task Force

H.323 Protocols

First of all to position the ITU-T's H.323 protocol architecture a brief overview is given for the H.32x standard series:

Recommendation	IP-based	Description
H.320	-	for narrowband switched digital ISDN
H.321	-	for broadband switched digital ISDN and ATM
H.322	-	for guaranteed bandwidth packet switched networks
H.323	X	for non-guaranteed bandwidth packet switched networks
H.324	-	for analog phone system (POTS)

It can be seen that H.323 is in alignment with best-effort Internet, since quality of service assurance is not expected from the underlying bearer services. [3.5.4]

ITU-T's H.323 protocol is generally called an umbrella protocol because it incorporates several sub-protocols (see Figure 3.5.1). Major protocols of H.323 protocol family are identified as follows:

H.225.0 CC – call and resource control protocol definitions.

H.225.0 RAS – registration, admission and status (RAS) control communication with the gatekeeper.

H.245 – defines the bearer control mechanism among H.323 terminals to establish audio and video connections.

T.120 – defines mechanisms for packaging and sending data.

H.450.x – definition of value added services.

H.323 Components

H.323 implementations distinguish four logical components, namely: terminals, gateways, gatekeepers and multipoint control units (MCU) (see Figure3.5.2). *Terminals* are the end-points of H.323 communications (data or signaling). Audio communication is mandatory while video or other multimedia service support is optional. *Gateways* as optional elements translate data formats, control signals; audio and video coding and provides call setup and termination functionality between different networks. *Gatekeepers* serve as the main control entities of H.323 zones. They ensure reliable communications and manage their zones centrally. Gatekeepers do not necessarily exist in all H.323 zones, though if one is present then all terminals must register with it. Gatekeeper's functionalities include address translation, admission and access control of end stations, bandwidth management

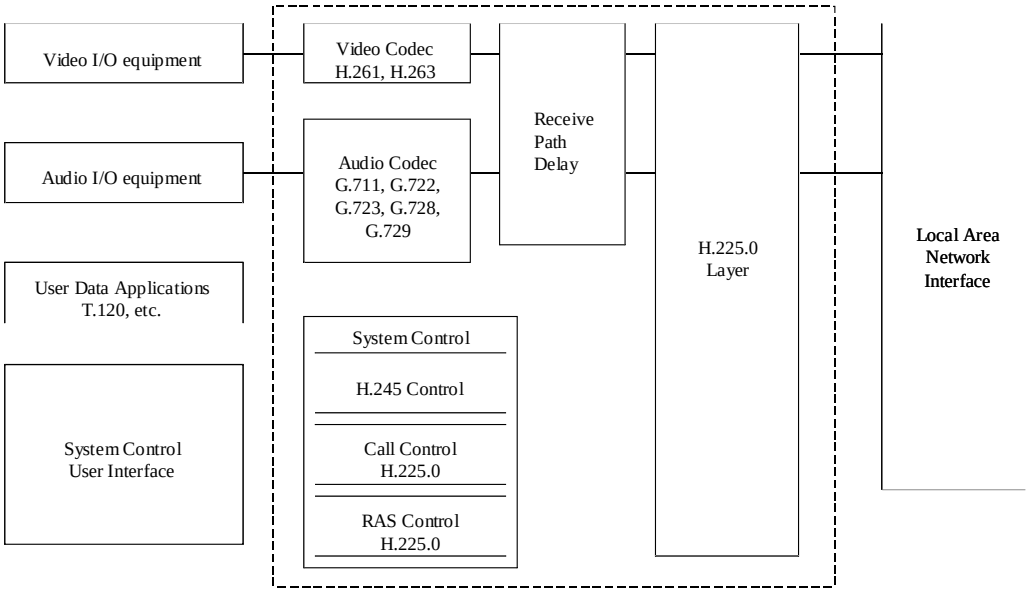


Figure 3.5.1 H.323 Protocol family

and routing capabilities [3.5.3]. Note that gatekeepers do not only serve conform but rather determine zone policy. *Multipoint control units* as optional element of an H.323 zone enable value added conferencing among multiple terminals. MCUs consist of two parts: i) a mandatory multipoint controller, which provides a centralized location for multipoint call setups and ii) optional number of multipoint processors (MP), which handle the mixing, switching and processing of audio and video streams among the terminals. Value added conference services that demand the presence of MCUs are converting among different media formats, presidential control of conference or multi site single picture conferences.

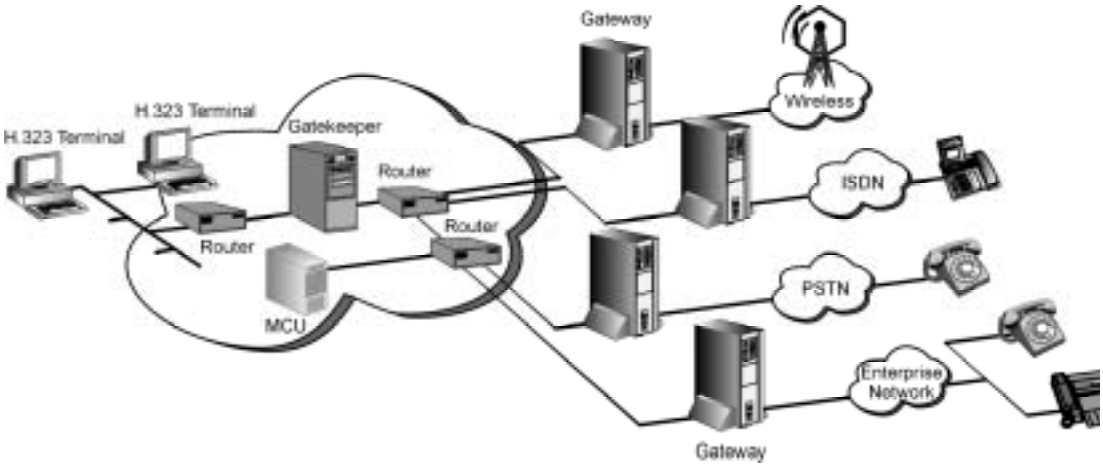


Figure 3.5.2 H.323 Components

H.323 Call Establishment

Even the simplest connection establishment method of H.323 reveals the high-level cohesion and dependencies within the protocol family. This makes the H.323 protocol architecture quite complex and monolithic. As an illustration Figure 3.5.3 shows the direct connection establishment between two H.323 terminals.

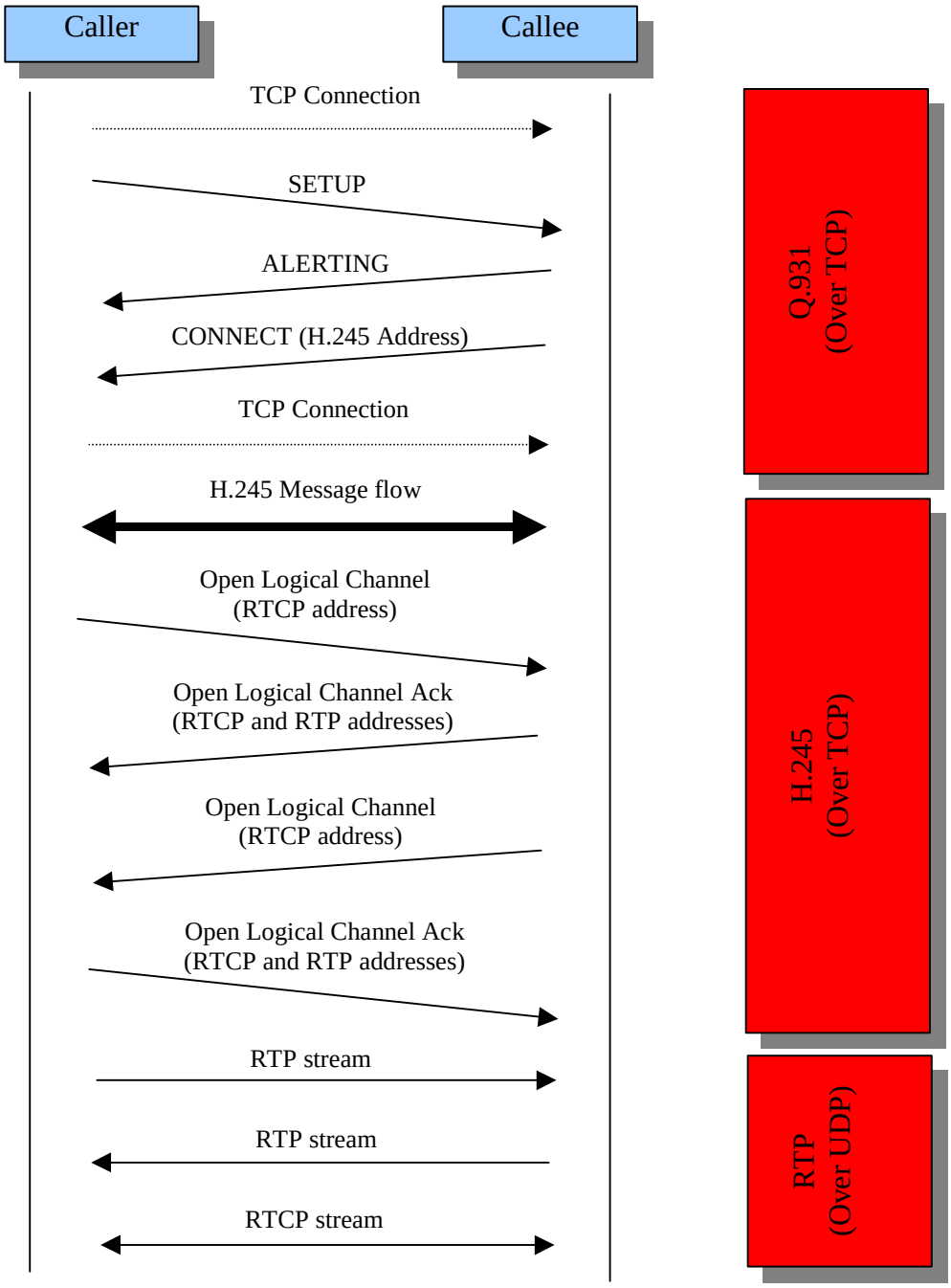


Figure 3.5.3 H.323 direct connection establishment

Note that the time that passes till the first bits of the streaming communication may be significant, i.e., several round trip times. In order to get by the long setup time; ‘Fast Connect’ has been introduced in H.323 version 2, where streaming media can start right after the ‘Setup’ and ‘Connect’ messages. Further improvements and supplementary services were added to the original protocol definition. Last release of the H.323 happened at the end of 2000; it’s current version is 4.

Session Initiation Protocol

SIP is specified by the IETF community and is aimed to be a text-based, simple request-response protocol with analogy to client-server architectures. Hence, requests are sent by clients and servers respond them. Most implementations contain both the client and the server capabilities. SIP requests can utilize any underlying protocols (TCP, STCP, UDP). Communication is established through intermediate elements (see bellow). However, once the path is discovered requests are exchanged directly between agents. The SIP protocol limits its scope to multimedia session establishment and termination and does not target any more functionality like resource management, data transfer or management of multicast groups. These functionalities are handled by other IETF standards; hence making SIP a really modular protocol. The basic functions of the SIP protocol includes: call forwarding, caller and callee party `number’ delivery, user mobility support, endpoint capability negotiation and authentication as well as managing dynamic membership of multicast conferences. [3.5.7]

SIP Components:

Similarly to the H.323 build-ups, SIP also consists of several entities that form a SIP network. *User agent client (UAC)* is a logical entity that initiates and sends SIP requests. This role lasts only for the duration of the transaction. *User agent server (UAS)* is a logical entity that receives and responds to SIP requests. It is generally done on behalf of the user. Requests may be accepted, rejected or redirected. This role also lasts only for the duration of the transaction. *User agent* is a logical entity that incorporates both UAC and UAS. *SIP Terminal* support real time two-way communication with another SIP entity; must include UAC. *Proxy* is an intermediate entity among SIP clients or next-hop servers; its function is to forward requests downstream and to pass responses upstream. There exist stateful and stateless

proxies. Stateless proxies solely forwards messages while stateful proxies do advanced signal processing. A *Redirect server* accepts a SIP request, maps the address into zero or more new addresses and returns these addresses to the client. Unlike a proxy server, it does not initiate its own SIP request. Unlike a user agent server, it does not accept calls. A *location service* provides information to redirect or proxy servers about the callee's possible locations. Location service may be co-located with other services (See Figure 3.5.4)[3.5.7].

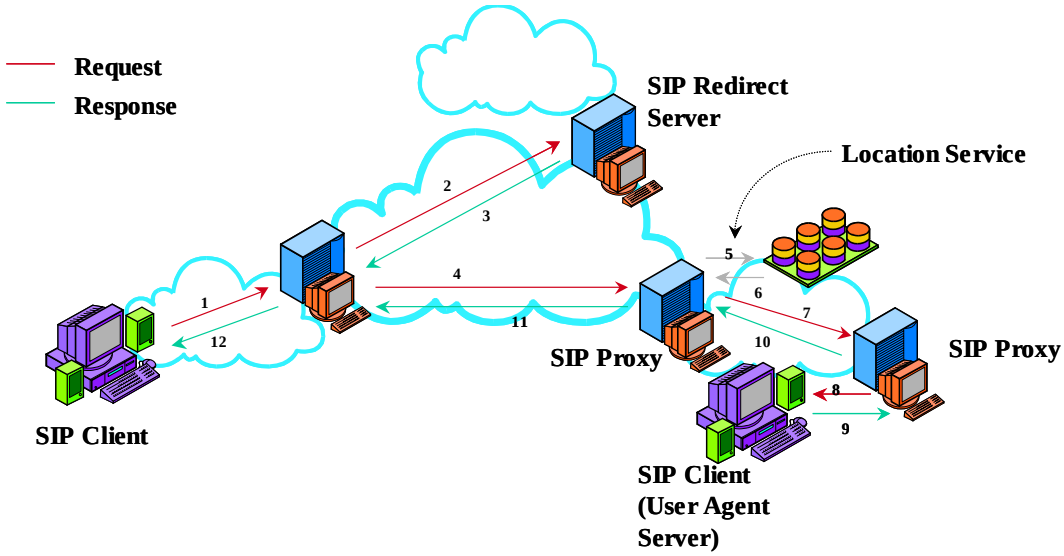


Figure 3.5.4 SIP Components and client location

SIP Call Establishment

SIP clients may initiate six major request types to the servers [3.5.7]:

INVITE	to inform the callee party about the parameters of the connection like the mediums and their port numbers
ACK	to confirm a request for call establishment; it may contain call parameters (see INVITE)
OPTIONS	to query the supported services of a server
REGISTER	to inform a server about user availability
BYE	to terminate a call
CANCEL	to terminate unsuccessful user lookups

Possible answers are also classified into six categories [3.5.7]:

1xx	Informational
2xx	Successful
3xx	Redirection
4xx	Request failure
5xx	Server failure
6xx	Global failure

Above figure (Figure 3.5.4) on the SIP protocol components illustrate a complex connection establishment. Protocol messages identified by the corresponding numbers are as follows:

Firs of all, the SIP client initiates the call (1-INVITE). The first intermediate entity is a proxy server, i.e., it acts on the behalf of the client in establishing the session. It generates a new (2-INVITE) message that arrives to redirect server; which in turn informs the requesting proxy of the new location for the user (3-3xx). The acting proxy is redirected; it sends a new (4-INVITE) message towards the new location of the user. Note the proxying function: it is not the client that is informed but the proxy that act on the behalf of the client. The next proxy server quires a location service (message 5 and 6) and becomes a client to further seek for the user (7-INVITE). Another proxy server with similar function is within the path (8-INVITE) when the callee is detected. An ACK is sent back along the proxy server path to the caller (messages: 9, 10, 11 and 12) by which the session is established.

Real-Time Protocol (RTP)

The main function of RTP is to carry real-time services, such as voice and video over an IP network, and so to complement the UDP/IP header. RTP provides payload type identification so that the receiver can determine the media type contained in the packet. Sequence numbers and timestamps are also provided for packet reordering, loss detection and playback speed adjustments. In addition RTP supports multicast communications. [3.5.6]

3.5.4. Comparison of H.323 and SIP

It can generally be said that both protocols are fast in adapting to new requirements and the early stage major differences between the protocols have long ago disappeared. Nevertheless, there are still the differences in the encoding of messages, i.e., H.323 uses binary message encoding (ASN.1 and Q.931) while SIP uses ascii (HTTPish) formalizations.

Both protocols may benefit from their standardization environment, i.e., H.323 may benefit form the high quality and very exact standardization process of the ITU-T; with lots of interoperability testing and well-developed gatewaying functions

towards POTS networks. SIP on the other hand may benefit the faster standardization process, increased flexibility and the resistance of the Internet community to external standards to IETF.

3.5.5. Future of IPT

Basically, any means of wide commercial deployment is delayed due to interoperability reasons and the lack of support for quality of service, charging, security, SS7 gatewaying etc…

On the other hand nowadays multimedia market is only IP telephony due to the supported low bit rate accesses (modem dial-ups). Since the technology is relatively new, time has to pass to have matured good products on the market. Then again companies can provide VoIP today at low cost and provide competition to the incumbent carriers. It again takes time to migrate the world from PSTN to IP networks. In the future, it is definitely video and other multimedia that users expect to see from the technical evolution. [3.5.8]

3.5.6. Summary

IP telephony is present; though there are some deficiencies that prevent its wide deployment. There are basically two remarkable standards competing for the market: ITU-T's H.323 and IETF's SIP. H.323 has reached better positions in current markets when compared to SIP, though it is said that the competition is not over yet. The introduction of the 3rd generation mobile services may push SIP forward in deployment.

Nevertheless, one of the most vital issues regarding IP telephony is quality of service support, since toll quality voice communication is supported in POTS. To fulfill the QoS requirements better coding mechanism and better bearer services must exist. However, current IP networks offer only very limited QoS capabilities, which hopefully will change in near future.

Another vital issue is interoperability; it affects heterogeneous IPT systems as well as interoperation with PSTN networks. There are server organizations investigating interoperability questions just to mention some: ETSI's TIPHON, iNOW!, IMTC etc…

Security may be mentioned as the third problem to be addressed in the future. ITU-T's H.323 offers methods for authentication, integrity, privacy and non-repudiation. SIP on the other hand uses achievements of secure HTTP transport development.

Overall, IP telephony seems to be the first step to real-time multimedia integrated services.

Referencies

[3.5.1] Bo Li, Mounir Hamdi, Dongyi Jiang, Xi-Ren Cao, Y. Thomas Hou "QoS-Enabled Voice Support in the Next-Generation Internet: Issues, Existing Approaches and Challenges" IEEE Communications Magazine, April 2000

[3.5.2] D. Bergmark, S. Keshav "Building Blocks for IP Telephony" IEEE Communications Magazine, April 2000

[3.5.3] Trillium, "H.323 Tutorial," 1999, 20 pages, http://www.webproforum.com/h323/

[3.5.4] Lasse Huovinen, Shuanghong Niu, "IP Telephony", http://www.tml.hut.fi/Opinnot/Tik-110.551/1999/papers/04IPTelephony/voip.html

[3.5.5] Communications Industry Researchers, Inc, http://www.cir-inc.com/

[3.5.6] H. Schulzrinne, S. Casner, R. Frederick, V. Jacobson "RTP: A Transport Protocol for Real-Time Applications", Internet Engineering Task Force, Internet Standard, RFC 1889, January, 1996

[3.5.7] M. Handley, H. Schulzrinne, E. Schooler, J. Rosenberg "SIP: Session Initiation Protocol", Internet Engineering Task Force, Internet Standard, RFC 2543, March, 1999

[3.5.8] H.323 forum at http://www.h323forum.org

[3.5.9] M. Handley, V. Jacobson "SDP: Session Description Protocol", Internet Engineering Task Force, Internet Standard, RFC 2327, April, 1998

3.6. ATM, IP over ATM, ATM-LAN and MPLS

Tibor Cinkler, author

Tamás Henk, reviewer

The term ATM is the abbreviation for **A**synchronous **T**ransfer **M**ode.

The purpose of ISDN network development was the re-utilization of the subscriber line, the most expensive and least utilized part of conventional telephone networks. However, this also brought about a disadvantage: ISDN is a narrow-band network. This is why it is frequently called also N-ISDN, i.e. narrow-band ISDN. (2 Mb/s is normally taken as the value separating the narrow-band and the wideband wireline access.) On the other hand, ISDN has the advantage of providing, in addition to telephone services, also data transmission, video conferencing, high-speed fax transmission yielding better quality, and value added services.

In order to meet the ever increasing demands, ITU-T begun to prepare the B-ISDN (Broadband ISDN) Recommendations. These were expected to refer to a broadband digital network providing integrated services, and in addition to point-to-point connections, also point-to-multipoint and multipoint-to-multipoint connections, and supporting both switched and permanent connections, further circuit switched and packet switched services, and both one-way and two-way (symmetrical or asymmetrical) connections.

ITU-T has defined a reference model for B-ISDN, a signalling system similar to that of ISDN, and found ATM to be the most suitable to meet the set of B-ISDN requirements.

3.6.1. Arguments for ATM

In data and telecommunication networks, traffic is distributed both in space and time. During day-time, trafic is high at places of work, but during periods after working hours and over week-end, traffic is shifted to residental districts. By combining the various traffics and transmitting them over a single network results in

more efficient traffic handling. Consequently, all data, telephone and program distribution networks have not to be planned for peak traffic conditions, and the traffic distribution in space and time allows then the sharing of resources.

Earlier, basically two types of networks have been utilized. Circuit switched networks were proven and widely used solutions for telephone networks, while packet switched networks prevailed in computer and data transmission networks.

In circuit switched networks, a connection (circuit) with a given capacity meeting the requirements and utilized only by the partners in question is established. As soon as the information transmission is finished, the circuit is disconnected, and the resources earlier utilized are made available for other connections. The quality of service (QoS), e.g. small delay and low data loss, can thus be guaranteed, provided the set-up of the connection was not confined by resouce capacity. However, on the other hand, the reserved channel of fixed bandwidth is then not exploited in a significant proportion of the time during burst-type data transmission, and it may even happen that the transmission capacity will be utilized in one direction only.

In the case of packet switching, the situation is reversed inasmuch as packets will only be transmitted if sufficient information has gathered. This can be achieved without expectation, but to insure the arrival of packets at their destination, the packets have to be ”addressed”. This means that a header is added to each packet. In this way, resources can be better utilized if sources are sending infrequently much information, but the network will provide no quality assurance.

Quality can only be assured if the required resources are allocated AND it is controlled wether the traffic sources are complying with the allocation. Therefore, ATM combines the above measures, and utilizes a so-called virtual circuit switching, fragmenting the information into small packets or cells with small fixed dimensions.

The deterministic multiplexing is thus substituted by statistical multiplexing. Presently, ATM is the only system providing an efficient utilization of resources, in addition to QoS guarantees. This holds for both the access network and the backbone network, in the 2 Gbit/s to 2.5 Gbit/s bitrate range.

The most important advantage of ATM, the quality of service QoS and simultaneously the efficient uitilization of resources, is paid for by system complexity (see Sec. 4.2). In IP networks, one of the most important trends at present is to

provide quality assurance, primarily because of real time applications, e.g. voice transmission over Internet Protocol (VoIP). This can be achieved only by increasing the complexity of the system by introducing the DiffServ and IntServ solutions.

3.6.2. Characteristics of ATM

A concise characterization of ATM is given in the following.

- ATM is the preferred transmission method of B-ISDN (Broadband ISDN).
- It is cell switched, data are transmitted in small packets of small dimensions.
- Virtual circuit switching is utilized, so that path selection is only once required, followed thereafter only by the simple transmission of the cells.
- Switching is realized by hardware, neither flow control nor correction is carried out at the nodes, only at the network boundary.
- By utilizing statistical multiplexing, arbitrary bandwidth can be reserved and utilized in the given bitrate range at the access terminals.
- Arbitrary quality can be provided with good resource utilization.

Let us now review the basic ATM technical concepts and solutions in order to shed light on the above features.

3.6.3. Technical background

What makes ATM to become an asynchronous transmission mode?

In the course of multiplexing, channels with differing bitrates can also be combined. At the output, an arbitrary number of free time slots can be utilized by the data units, i.e. they are not synchronized to a single framing. But in order to facilitate the assortment of these units at the receive side, they are all labeled.

What is the composition of a cell, the unit of data transmission?

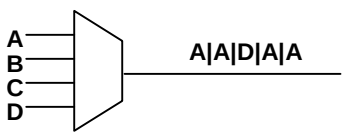


Figure 3.6.1. Asynchronous

The composition of a cell on a Network-to-Network interface (NNI) is shown in Figure 3.6.2. Each row represents an octet, and the contents of these rows are read out line-by-line, from left-to right, going downwards from the top row. The first 12 bits represent a VPI (Virtual Path Identifier), showing the VP (virtual path) to which the given cell is attached. The next 16 bits represent the Virtuial Channel Identifier (VCI) which has the function of identifying the connection within the group denoted by the VPI value.

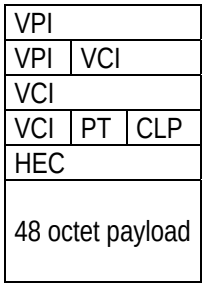


Figure 3.6.2. Composition of an ATM cell

The 3-bit Payload Type (PT) field is intended for identifying the payload and controlling congestion. Finally, the Cell Loss Priority (CLP) field shows whether the given cell can be thrown away with high or low priority, in the case of network congestion.The Header Error Control (HEC) has two functions, either header error detection and repair, or detection of the cell boundaries. The header with 5 octets is followed by 48 octets of relevant data, so the complete cell comprises 53 octets.

If not a network interface but an interface between the user and the network (UNI) is investigated then the first 4 bits of the cell represent the Generic Flow Control field (GFC) that has been designed for traffic control purposes.

How does ATM fit into the ISO OSI reference model?

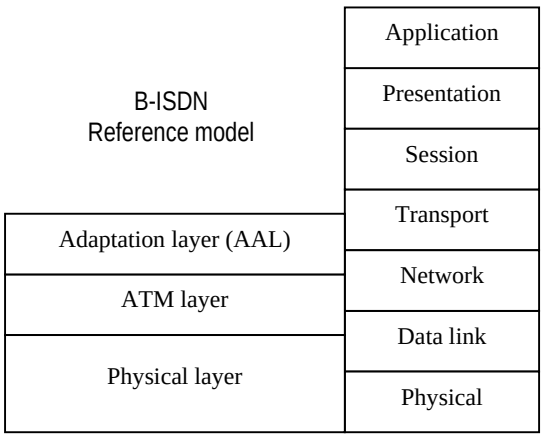


Figure. 3.6.3. Comparison of the OSI and B-ISDN reference models

ATM corresponds to the B-ISDN reference model as defined by ITU-T. This is a rather complicated three-dimensional model but the Figure shows only one dimension. For ATM, ITU-T has defined only three layers. Note that compared to the OSI reference model, the physical layer of ATM is more sophisticated, and this deviation, as shown in Figure 3.6.3 is also present at the other layers. There are several kinds os adaptation layers, the most wide-spread being the AAL5 (ATM Adaptation) layer which is very simple, has only few functions, its main function being the fragmentation of the higher level data units during transmission, and their de-fragmentation at reception, respectively. The IP, FR and PDH) traffic signals are transferred to the ATM network by the adaptation layer.

What sort of addressing is applied in ATM networks?

In public ATM networks, Rec. ITU-T E.164, used in telephone networks, is applied. The addressing in private networks is different. (Figure 3.6.4)

What sort of „links” are present in ATM networks?

A link carrying data traffic between two terminals is a ”virtual circuit” or ”virtual channel circuit” (VCC). This VCC is realized over several sections by utilizing several

Figure 3.6.4. ATM addressing

VPs (virtual paths).

A virtual channel circuit can be either permanent (PVC) or switched (SVC). A PVC is established by means of configuration by the network management system or the network element management system while the SVC is initiated by the subscriber over the signalling system.

In addition to the point-to-point connections, ATM supports the point-to-multipoint connections too. Even connection-less applications can be realized over ATM connections. (Figure3.6.5)

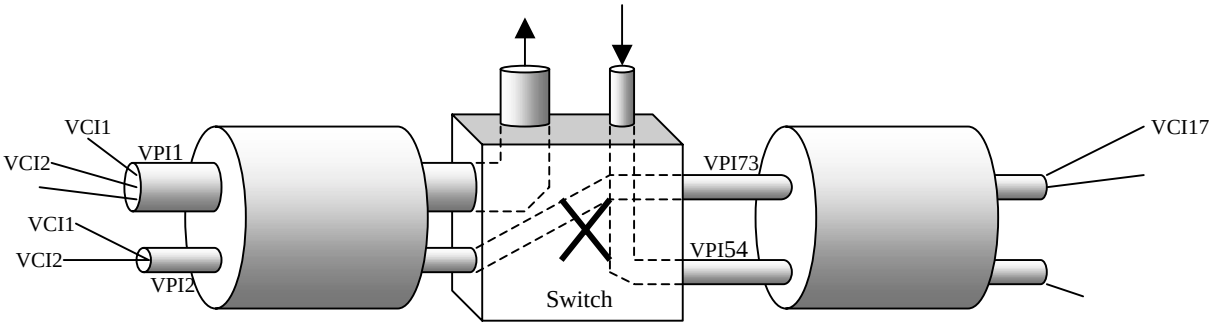


Figure 3.6.5. VP and VC switching or cross connecting

Traffic management

ATM networks have excellent traffic management. As already mentioned, ATM is the only solution capable of providing good resource utilization with arbitrary quality assurance.

As given by ATM traffic management specifications, sources are classified to fall into several groups according to the *traffic* and *quality* requirements. These are only listed in the following; for more details, see Ch. 4.

- CBR: Constant Bit Rate: Characterized by constant bit rate, see e.g. the leased line
- rt-VBR: Variable Bit Rate, real time source with variable bit rate
- nrt-VBR: non-real time VBR
- ABR: Available Bit Rate: source with available bit rate, with the transmit bit rate controlled by feedback, depending on the traffic load of the network.
- UBR: Unspecified Bit Rate. Sourcee with unspecified quality parameters.

- GFR: Guaranted Frame Rate: A source requiring the transmission of larger bit rates. Here the relations between frames fragmented into cells are registered.

The reservation of resources is based on traffic and quality parameters, their fulfilment being controlled over the total connection time by the user and the network, respectively.

3.6.4. How can a computer network be established over ATM?

This can be performed in several ways.

1. Multi-protocol encapsulation (IETF RFC 1483)
2. Classical IP over ATM, CLIP (IETF RFC 1577)
3. LAN Emulation over ATM Version 1.0 (ATM Forum LANE 1.0)
4. ATM Application Programming Interface (ATM API)
5. Classical IP and ARP over ATM (IETF RFC 2225)
6. Multi-Protocol over ATM) (ATM Forum MPOA)
7. LAN emulation, 2nd version (ATM Forum LANE 2.0)
8. MultiProtocol Label Switching (IETF MPLS)

These techniques can be classified according to the purpose to be achieved. Methods 3, 6 and 7 have been devised mainly for the transmission of LAN frames while methods 2, 5 and 8 are intended for the direct transmission of IP packets. Method 1 is suitable for the transmission of both LAN and IP frames. Accordingly, LAN type solutions are suitable for spanning smaller distances, subject to media access, while solutions resulting in the transmission of direct IP packets can be applied for spanning larger distances.

3.6.5. Multiprotocol encapsulation (IETF RFC 1483)

This was the first solution enabling IP based communication over ATM networks. This solution was elaborated by IETF, its basic concept is contained in RFC 1483, adopted in July 1993. The attribute *multi-protocol* in the designation is justified by the fact that a signle PVC is shared by several LAN frames and protocols. A single encapsulation is used only, notably Link Layer Control (LLC) encapsulation. In the solution introduced in RFC 1483, only PVCs, i.e. permanent virtual channels are applied. Thus permanent channels are established for all parties participating in the communication, independently from the fact whether they transmit data or not.

These PVCs are termed tunnels as these are only pipes transmitting bits. In this network, two workstations communicate typically over bridges or routers in order to avoid the full PVC mesh, alltlhough they would be able to communicate directly. A tunnel is intended to connect either birdges or routers, but not simultaneously. (Figure3.6.6)

One of the drawbacks of this solution is the permanent reservation of the resources by the PCs (PVCs), including the VPI/VCI address space. Also, the path selection capabilities of the ATM VC-switch (SVC) are neglected by this solution.

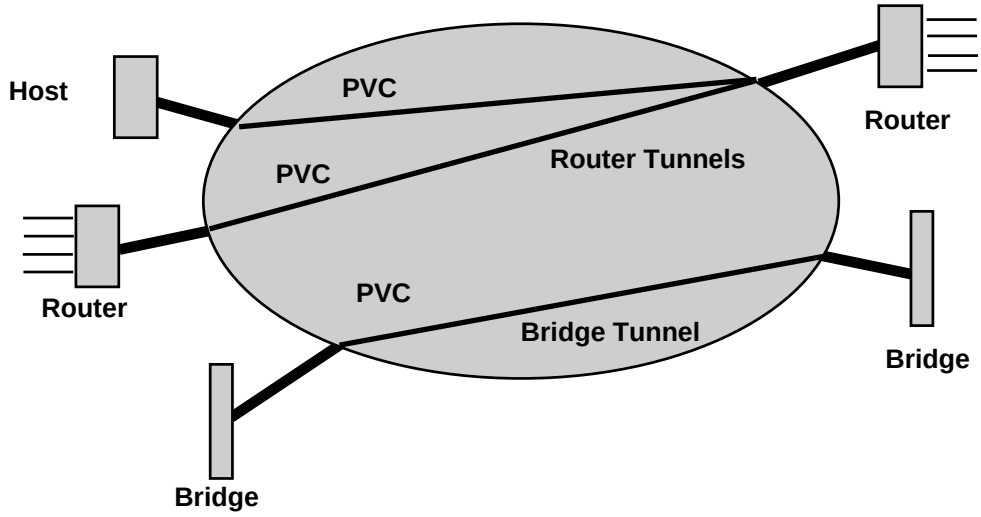


Figure 3.6.6. "Tunnels" in multi-protocol encapsulation

3.6.6. Conventional IP over ATM (IETF RFC 1577 and RFC 2225)

This is a more complex, but also more flexible solution, capable of utilizing the advantages of switched VCs (SVCs). However, it has the disadvantage of supporting only the IP, and not supporting other network protocols. As seen from its designation, it comprises the transmission of TCP/IP packets over the ATM network and the address resolution. Its basic concept is defined by RFC 1577 and 2225, as specified by IETF (Internet Engineering Task Force).

Referring to the IP packets, their fragmentation at the transmit side, and their de-fragmentation at the receive side are carried out by the AAL5 Adaptation Layer. This method supports both PVC and SVC based connections.

Workstations that have to set up direct connections with each other are classified as Logical IP Subnetworks (LIS). Three LISs of this kind are illustrated in Figure 3.6.7. Workstations within a LIS are capable of seting up a direct ATM link between each other but those within different subnets can only reach each other over several SVCs by utilizing routers.

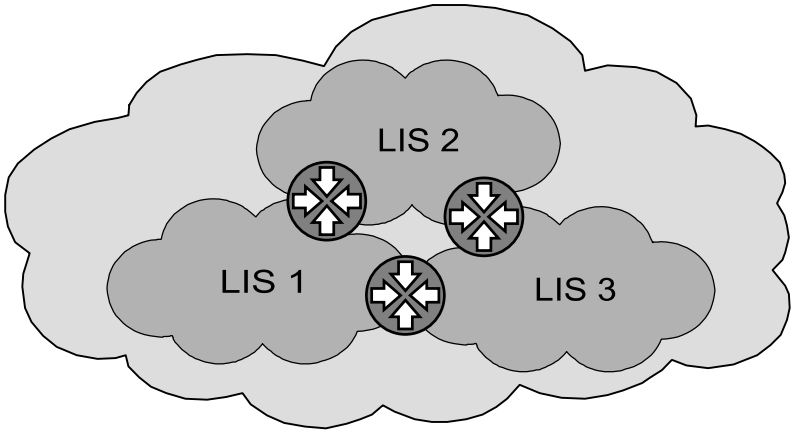


Figure 3.6.7.Logic IP subnetworks and their connections

In the following, the establishment of connections within the individual LISs will be investigated. All IP workstations to be connected to the network over the ATM should have an ATM address. When a PVC is applied (e.g. when the network does not support a signalling system) then each workstation has its own address resolution table, i.e. to each IP address, a given VPI/VCI label is assigned. If SVC is applied, then each LIS should have an address resolution server (ARP: Address Resolution Protocol).

Let us now investigate the set-up of an IP session.

First of all, the client will establish a connection to the ARP server which ”runs” either over a separate workstation or, mor frequently, over an ATM switch. The ARP server address can be configured or dynamically asked for. Subsequently, the ARP server will send an inverse ARP request to obtain also the IP address corresponding to the ATM address of the connected workstation. The address resolution table is thus kept up-to-date by the ARP server.

Assume now that the workstation A1 would like to reach workstation A2 by utilizing its IP address. First, workstation A1 tries to look it up in its own address resolution table. If not found there, workstation A1 requests from the ARP server the

A2 ATM address, and subsequently establishes an ATM SVC connection to A2. This is followed by data transmission, and following an extended time without data transmission, the SVC will be disconnected. Upon a repeated attempt of A1 to reach A2, its ARM address will this time be found in its local address resolution table. In this way, the connection build-up time is accelerated, and the loading of the ARP server is decreased.

If A1 would like to reach a workstation (e.g. B1) that is not located in the same LIS then this can be achieved only over a router. In Figure 3.6.8, the ARP and the router coincides but this is not necessarily the practical case.

The point-to-multipoint connection has not been supported by the conventional IP over ATM. This function has been realized by letting each node to establilsh a point-to-multipoint connection towards all others. Compared to this solution, the MultiCast Server (MCS) requires less resouces, inasmuch as the message reaching the server from any node will be multicast to each terminal. This is realized by assigning several ATM addresses to a single IP multicast group address.

RFC 2225 recommends the use of several ARP servers, with their data bases permanently synchronized to each other. The network will thus be less prone to failures.

The advantage of the conventional IP over ATM solution is the capability of the IP network to utilize the quality assurances provided by ATM.

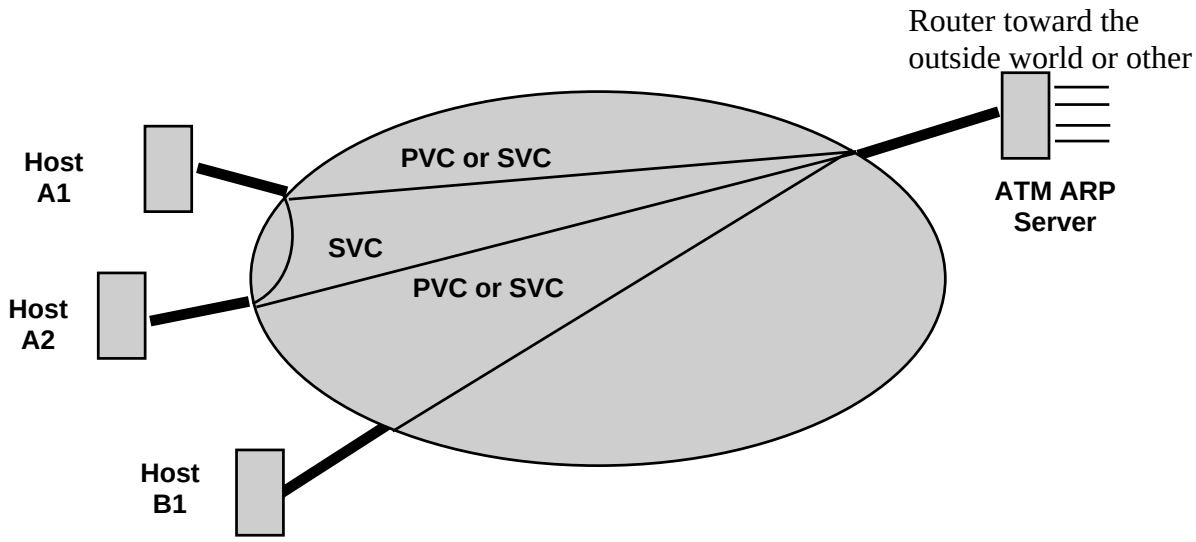


Figure 3.6.8 Conventional IP over ATM: oneLIS

3.6.7. LAN emulation (ATM Forum LANE 1.0 and LANE 2.0)

The LANE emuation, as opposed to the CLIP solution, is not only intended to transmit the IP traffic but also to emulate LAN (e.g. Ethernet or Token Ring), as suggested by its designation. Thus all protocols in the third layer can be used without change, e.g. IPX, DecNet, Novell Netware Client, NetBEVI, etc.

Over an ATM network, emulated or virtual LANs (ELANs) can be established. To these, workstations with ATM interface cards or, via bridges with ATM interfaces, complete subnetworks can be connected. Here too, workstations within a given ELAN can directly communicate with each other via SVCs. (Figure 3.6.9) However, stations in different ELANs can only communicate over routers, where the interconnection between routers itself is also an ATM connection over ATM switches/cross-connects However**,** another disadvantage of this solution is the lack of quality assurance because the LAN which is emulated is not capable itself to provide this.

The realization of the LAN emulation requires clients (LEC: LAN Emulation Client) and a server consisting of three modules as follows:

- **LECS:** LANE Configuration Server
- **LES:** LANE Server; its main function being the registration of members (identification, control functions), and address resolution.
- **BUS:** Broadcast and Unknown Server, having an unknown address and serving broadcast transmission. This is something like the MCS with the CLIP.

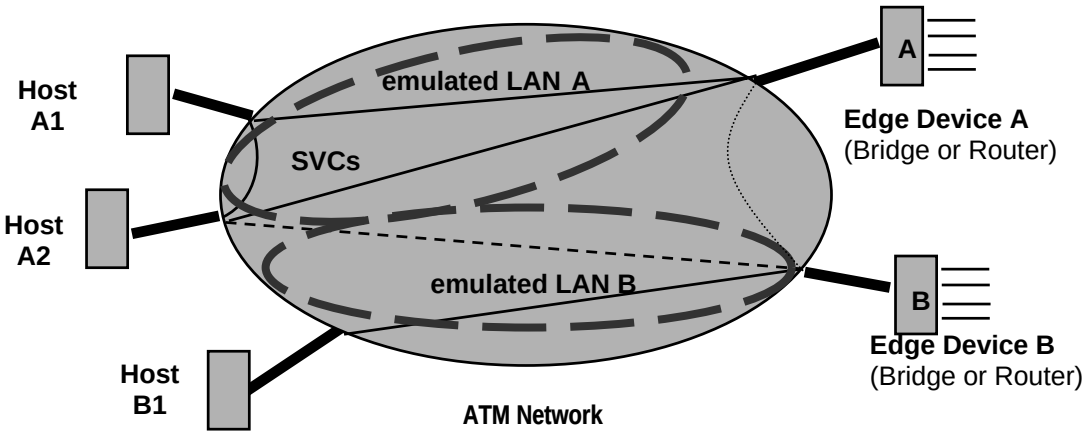


Figure 3.6.9.Emulated LANs and their relation in LANE

Its main function is the realization of point-to-multipoint broadcasting. If a message with unknown address is sent, this server emulates the shared medium, and communication is established over this link too until direct VCC is established between the two terminal stations.

The interface between the servers providing the LEC and LANE services is the LAN Emulation Unit (LUNI).

Let us now investigate the operation of the LANE. Data transmission is preceded by putting into operation, and the establishment of connections required by the data transmission, by means of signalling messages.

The LEC first asks the LECS to provide configuration information. To this end, first a temporary "Configuration Direct VCC" is established, followed by the establishment of a ”Control Direct VCC” intended for address registration and address assignment towards the LES. This is foillowed by establishing a unidirectional ”Multicast Send VCC” towards the BUS, whereupon the BUS establishes a multicast forward or distribute unidirectional point-to-multipoint connection towards the LECs. (Figure 3.6.10)

Subsequently, when a LEC sends a message/packet over the LUNI, a table is first scrutinized whether a VCC is already existing towards the given Media Access Control (MAC) address. If so, transmission takes place to the MAC, if not, an ATM address is required from the LES, based on the given MAC address. Thereafter, a ”Data Direct VCC” is set up. While this VCC is set up, packets can be sent over the

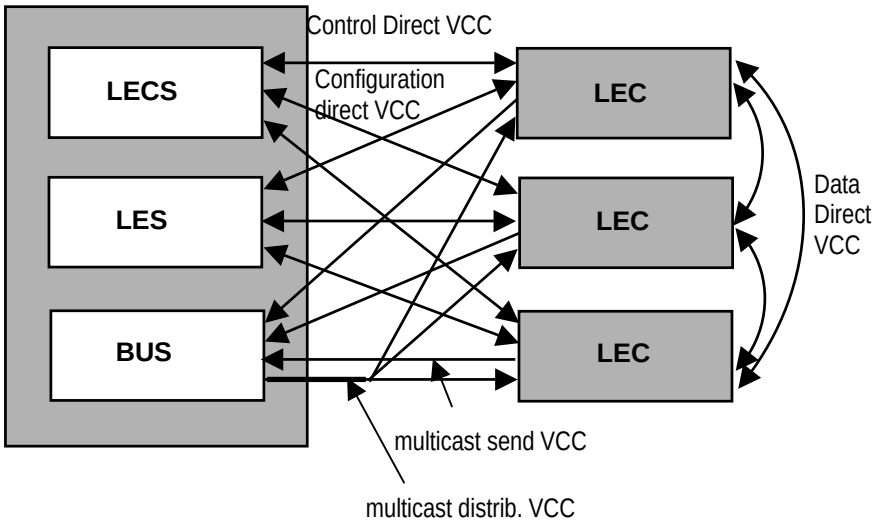


Figure 3.6.10. The LANE clients and elements of the server mulating the LANE

BUS too. The unused connections ”expire” and will be disconnected. The unused MAC address expires too, and if no suitable LEC is found, it will expire too. This renders possible the movement of LECs in an ATM network.

Note that here we have a double address assignment: the conventional IP address – MAC address assignment, and the MAC address – ATM address assignment. Only the latter is the task of the LANE.

The protocol stack shown in Figure 3.6.11 illustrates the method by which a connection to an ELAN is established. At left, the connection of a workstation having an ATM network card is shown, while at right, the connection of a LAN (e.g. Ethernet) network or a workstation having a LAN card is presented.

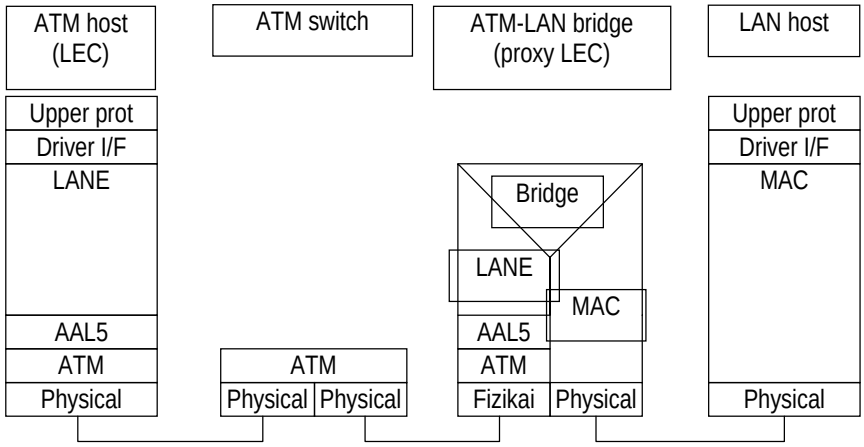


Figure 3.6.11. Protocol stack of the LANE with proxy-LEC

The LANE utilizes the AAL5 adaptation layer too. The proxi LEC renders possible the ELAN connection of LAN workstations or complete LAN subnetworks.

The LANE 2.0 version includes a few additional features: several servers can be applied within a single ELAN, the upper layers can access the QoS features provided by the ATM, the multicast communication is improved, several upper layer sessions can share a single VCC, and the MPOA solution is also supported.

3.6.8. ATM API

The use of an ATM application programming interface is justified when an application over the own ATM has to be established, or the existing applications have to be improved. The manufacturers of ATM network interfaces supply function libraries with their cards in order to support the development. This solution requires

both knowledge and time. As an example, the connection of workstations applying a Linux operational system to ATM based networks is supported by the Linux-ATM project.

3.6.9. Multiprotocol over ATM (ATM Forum MPOA)

In the case of LANE, clients in separate ELANs cannot set up a direct connection, but MPOA is able to provide this possibility. Originally, MPOA was started by the initiative of Newbridge, but has partly been incorporated into LANE 2.0.

Figure 3.6.12 illustrates the steps taken to establish direct connection between workstations within separate ELANs. First, connections are established over several MPOA servers (MPS). Subsequently, upon getting back the ATM address corresponding to the IP address, a VCC shortcut is established. This task requires the extension of the address resolution by using the NHRP (Next Hop Routing Protocol).

Similarly to LANE, MPOA is also based on a client-server architecture. The system is backward compatible with LANE: the MPC (MPOA client) includes also the LEC functions while the MPS (MPOA server) has the functions providing LANE services.

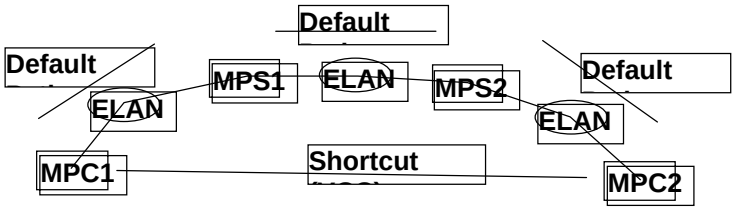


Figure 3.6.12. Set-up of connection in MPOA

3.6.10. Multiprotocol label switching (IETF MPLS)

Several methods for setting up an IP network over ATM have been dealt with. However, the features of simple solutions are restricted while sophisticated methods are far too complicated. The role of ATM has been essentially limited to the transport of the IP traffic so following question comes up: is it expedient to harmonize two routing methods, two addressing methods and differing signalling systems, with unavoidable complications, or wouldn’t is be better to combine these techniques?

The decision favoured the combination concept. In to-day's MPLS implementations, typically ATM hardware is utilized, and the labeled cell-type information transmission over virtual circuits has been preserved (this is called Label Switched Path or LSP). However, the routing and addressing are based on the solution used in IP networks.

The benefit of MPLS, as compared to conventional IP networks, is the more homogeneous traffic distribution over the network due to the LSPs, resulting in better network utilization. Dynamic traffic management is possible, allowing traffic direction, resource reservation and quality assurance. Further, the routing decision for each single packet is omitted, this being necessary only once per stream, and thereafter, label based packet transmission is applied.

Considering precedents, note the Cell Switching Router, proposed by Toshiba in 1994, or the IP switching solution, proposed by company Y in 1996. Also in 1996, IBM proposed the ARIS solution (Aggregated Route Based IP Switching). Finally, the nearest resemblance to today's MPLS system shows the Tag Switching system proposed by CISCO in 1996. All solutions are based on the preceding concepts.

The MPLS solutions have several development stages. Earlier, the LSPs have been fixed, but in course of the development, they could be dynamically established and disconnected according to demand, with a traffic management similar to ATM. MPLS is able to handle not only ATM but also other techniques such as Ethernet, Frame Relay, Token Ring and others. However, the ATM based solution is most widely used. It is called multi protocol because over MPLS, not only IP but more generally, also voice, video, multimedia and data transmission applications can equally be realized.

The MPLS principle has been generalized for wavelength routing (WR) DWDM networks, based on the wavelength information instead of the label. This is the MPLambdaS network where Lambda refers to the wavelength information. This way, the combination MPLS/MPLamdaS allows the setting up of more efficient two-layer networks. These can further be generalized to obtain GMPLS, with 4 to 5 network layers stacked on each other. GMPLS is the abbreviation of generalized MPLS, referring to the method according to which a waveband switched layer or directly a wavelength switched layer is stacked on a fibre switched layer. This is followed by a TDM layer with larger time frames (e.g. with SDH-like framing), and on this is stacked the packet-switched or cell-switched upper layer.

Translated by Tamás Sárkány dr.

3.7. Quality of Service in IP Networks

Róbert Szabó, author

György Réthy, reviewer

3.7.1. Definition

The exact definition of quality of service (QoS) is difficult to give, however it can be seen as the *ability* of the service provider (layer, network entity, etc…) *to offer* as a service to consumers (upper layers, users, peer-entities, etc…) throughout which certain requirements (probabilistic or deterministic) can be met as long as the traffic conforms to the negotiated traffic contract. The former definition incorporates network layers' QoS as well as peer-to-peer edge-to-edge or end-to-end QoS deliveries. Naturally, the requirements may vary from application to application or from the point of the delivered level of QoS.

3.7.2. Overview

Traditional Internet offered the very same level of quality of service (QoS) to the online community, which was named best-effort service. Most differentiation among customers has been probably only in the connectivity type: analogue modem vs. ISDN, or dial-up vs. leased-line [3.7.3]. In the recent years though, ISPs increasingly demand new ways of service deliveries in alignment with new emerging applications like multimedia services or virtual private networking. Furthermore, the Internet has to face the converging world where the early-days slogan of "IP over everything" seems to change to "Everything over IP". This again stretches the inventions of IP that is dated back to the 60s'.

In harmony with the emerging demands the last decade of Internet related research has dedicated a substantial effort to seek for methods, protocols and architectures of the future's Internet.

This section is organized as follows: Application level QoS requirements are categorized first and then different approaches to deliver QoS in the Internet are summarized.

3.7.3. Application Requirements

Applications can be regarded as the driving force of network developments. In another view, it is the network that exists to serve application demands and it should not be the application that renders itself dull due to network inadequacy.

Applications can be characterized along two attributes: i) their predicted transmission rates and ii) their delay tolerance [3.7.1]. According to transmission rate characteristics, applications may be classified into the following categories:

Stream	Predictable and relatively constant transmission rate (constant bit rate – CBR). Applications in this category have explicit upper bounds on their rates (peak rate). Examples are video streaming, voice over IP etc…
Burst	Unpredictable and varying transmission of *blocked* data. There is no explicit upper bound on their peak rates. Example is file transfer.

Table 3.7.3.1 Characterisation of applications

According to their delay tolerance, applications may fall into the following categories:

Asynchronous	Elastic application without any timing demands. (E.g.: mail transport, file transfer)
Synchronous	Flexible applications with loose timing requirements. (E.g.: web browsing)
Interactive	Applications where usability and functionally do not depend on timely delivery but user satisfaction at some extent. (E.g.: telnet, web browsing)
Isochronous or conversational	Usability is affected by the perceived delay characteristics. (E.g.: IP Telephony)
Mission-critical	Functionality is spoiled if delay requirements are not met. (E.g.: toll quality IPT)

Table 3.7.3.2 Requirements necessary of different applications in critical cases

3.7.4. QoS Frameworks

Networking solutions have evolved to satisfy the different needs of the application described formerly. First it was only the bandwidth that played the most important provisioning role in the Internet. By providing more and more bandwidth any of the application demands can be satisfied. This solution is called over-provisioning of network resources. However, since it is generally believed that there is no way of providing unlimited resource all through the network and as application

characteristics and requirements have changed during time intelligent methods were researched to adapt the network to the needs of new applications. It was the *integrated services* approach that first appeared with high level analogy to ATM networks' resource provisioning. This solution, in alignment with ATM, has turned to be computationally far too complex to be widely deployed; especially in core networks. By learning of the failures of the integrated services approach, nowadays research is focusing on a scalable solution that provides resource prioritization in contrast to resource reservation methods. This effort is hallmarked with the *lightweight* attribute and is named *differentiated services*.

Another key factors to the success of QoS Internet are the interpretabilities, proper policy managements, authentications and accounting/billings. These later functions are out of the scope of this study, while integrated services and differentiated services approaches are detailed in the following sections.

3.7.5. Integrated Services

In this subsection the detailed architecture of integrated services (IS) framework is described.

Bearer Services

In the IS framework the IETF has defined two additional bearer services to the traditional best-effort service, namely the *Guaranteed Quality of Service* [3.7.10] and the *Controlled Load Quality of Service* [3.7.11]. Both of the defined services provide controlled delays and packet losses, assume resource reservations and call admission control mechanisms in the background and use token bucket traffic descriptors [3.7.12]. However, there are fundamental differences between the two bearer services:

The *Guaranteed Quality of Service* provides strict upper bound on the maximum delay and zero packet loss for connection that conform to their traffic profile [3.7.10]. Note that this service neither guarantees the service's minimum or mean delay nor the volume of the delay jitter. It solely gives a deterministic upper bound on the maximum delay, which is guaranteed through bandwidth allocation. Unfortunately, the given delay bound is in most cases overly pessimistic in the means

that the experienced mean delay is far less than the negotiated upper bound [3.7.11]. The guaranteed quality of service network element requires strict call admission control that is performed on the user provided token bucket (also called leaky-bucket) descriptors.

In order to offer guaranteed QoS service all intermediate network elements along the communication paths must support this service. Nevertheless, partial support of this service can as well improve network performance, e.g. when congested links support while over-provisioned links do not.

The *Controlled Load Network Element* service offers a virtual service equivalent of the one an unloaded network would offer. Note that the definition does not contain any exact specification on the service quality; hence the specification is very simple. This simplicity was the interest of IETF [3.7.11]. An application requesting controlled load service can expect i) virtually no loss and ii) delays around the minimal network delay, i.e., very low delay jitter. Similarly to the guaranteed quality of service strict call admission control is necessary. The base for the admission control is the previously mentioned leaky-bucket traffic descriptor.

Traffic Policing

According to the admission control, if a connection is admitted into the system and its traffic is within the negotiated traffic volume and characteristics, the network must provide the negotiated QoS. Network boundary routers perform the supervision of user traffic conformance as well as authentication of users for resource reservations. These functions are called *traffic policing*.

The most common objective of traffic policing is to limit the maximum rate of traffic transmitted or received on an interface, i.e., to establish *bandwidth management through rate limiting*. This policy is mostly enforced at the edges of a network. Traffic that falls within the rate parameters is transmitted, whereas traffic that exceeds the parameters is dropped or transmitted with a different priority.

The most common traffic regulator is the so-called token or leaky bucket shaper (see Figure 3.7.1). The token bucket is described with a parameter-triple, namely the token rate (r), the token bucket length (b) and the peak rate (p). The peak rate is often ignored if p>>r, when is assumed to be infinite. Now, one can either use

a token bucket to regulate (shape) the traffic according to the parameters or to drop, mark or re-prioritize the traffic share above the contract. For a token bucket constrained arrival see Figure 3.7.2.

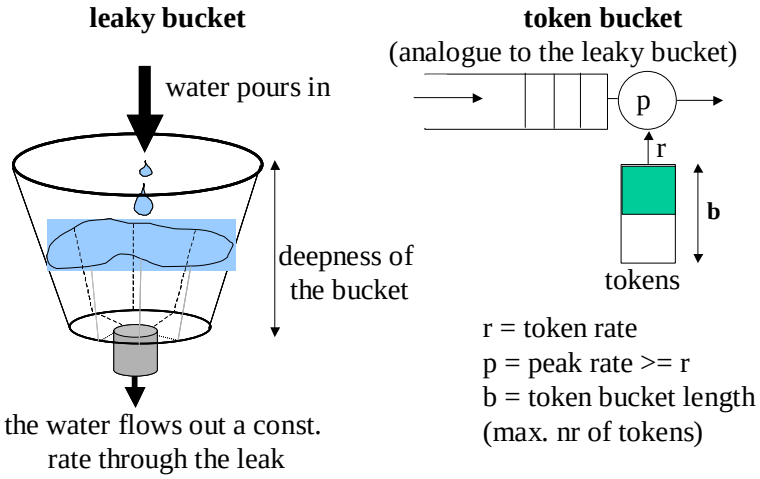


Figure 3.7.1. Leaky and token bucket

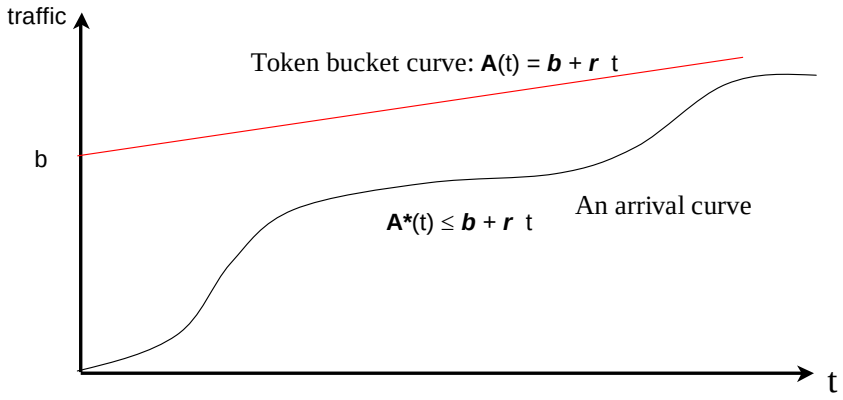


Figure 3.7.2. Token bucket constrained arrivals

Traffic Control

In the network element, the method through which QoS is provided is named traffic control. Traffic control incorporates i) packet scheduling, ii) packet classification and marking and iii) the previously described admission control functions [3.7.13] (see Figure 3.7.3).

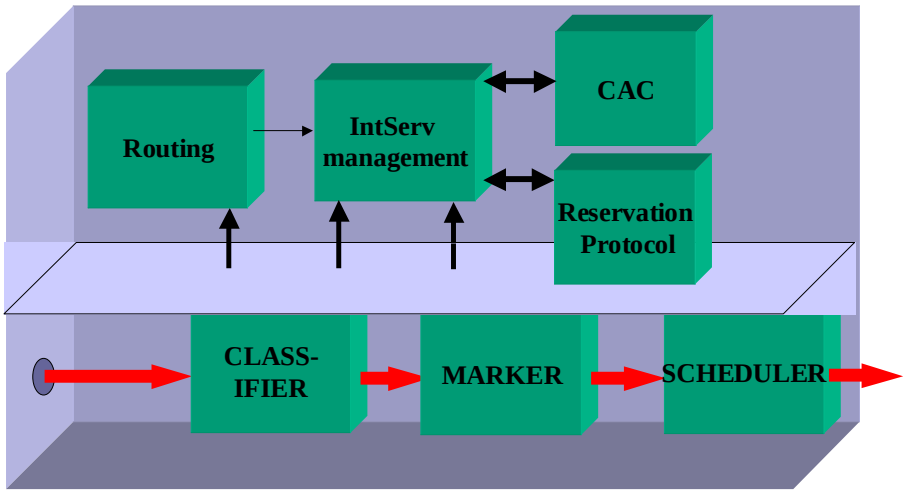


Figure 3.7.3. An IntServ node

Scheduling

Packet scheduling plays an important role in all QoS networks, since it provides the method for preemption among competing packets. The ideal packet-scheduling paradigm that gives the basics of all integrated services networks is the *weighted fair queueing* (WFQ) scheduling, which is a packetized derivative of the generalized processor sharing (GPS) scheduler. GPS schedulers ensure high level of traffic separation, rate-based resource allocation and re-distribution of excess system capacity. Furthermore, for any given traffic scenario described by token bucket parameters (r_i,b_i) i=1,..,N and server capacity of C, each of the traffic streams are ensured of a worst case delay bound ($D_i=b_i*\Sigma r_j/C*r_i$; i=1,..,N) and of zero loss if their allocated buffers are $B_i>b_i$; i=1,..,N and the system is stable, i.e., $\Sigma r_j<C$ (see Figure 3.7.4). In a common scenario one weight is assigned to each of the individual queues (Φ), which determines the bandwidth share of the connection.

It is important to understand, that IntServ in general requires per flow level buffering[3] within the nodes (see Figure 3.7.4.) hence has bad scaling properties when considered in backbone networks.

Resource ReSerVation Protocol (RSVP)

The resource reservation protocol (RSVP) is an end-to-end signaling protocol for resource management [3.7.14]. It can handle individual or filter defined group of

[3] Each session or connection has a separated queue hence inter-disturbance is minimized.

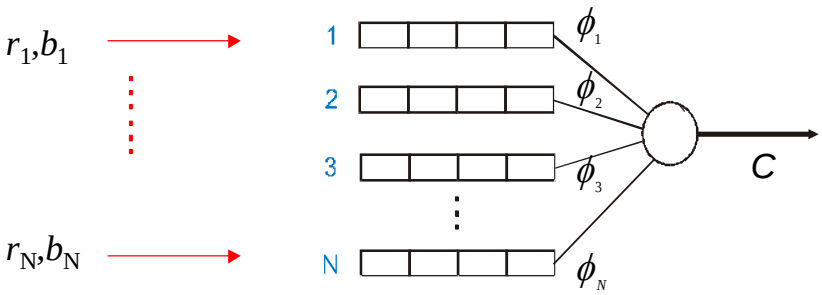


Figure 3.7.4. Weighted fair queueing scheduler

flows; establishes one-way resource reservation originated from the receiver side; has multicast support and efficiently merges reservation request for in multicast distribution trees. Although RSVP creates one-way reservation, both parties can initiate it in parallel and independently. RSVP typically resides in the transport layer when described, however it does not deliver application data but manages applications' resources. Sometimes RSVP is assumed to control communication paths as well, though RSVP is independent of the used routing protocols, hence it is transparent to varying routing along the paths.

RSVP maintains the reservation according to a soft-state paradigm, i.e., it periodically updates reservation states. This is in alignment with IP's connectionless approach. If for any reason reservation updates are cancelled, resources are automatically released in network elements. Naturally, there is also an explicit signaling message to tear down explicit reservation in order to not to wait until timeouts.

3.7.6. Differentiated Services

The DiffServ architecture is based on a simple model where traffic entering a network is classified and possibly conditioned at the boundaries of the network, and assigned to different *behavior aggregates* (BAs), with each BA being identified by a single DiffServ Code-Point (DSCP). Within the core of the network, packets are forwarded according to a *per-hop behavior* (PHB) associated with the DSCP. The smallest autonomic unit of DiffServ is called a DiffServ domain, where services are assured by identical principles. A domain consists of two types of nodes: boundary routers and core routers. Core nodes only forward packets, they do not participate in signaling. This architecture achieves scalability by implementing complex

classification and conditioning functions only at network boundary nodes, and core routers store no information about individual flows. The DS field in every packet's header stores the necessary information to identify proper queueing, scheduling and dropping mechanisms, which are called *per-hop behaviors* (PHB). Although it is the PHBs who determine QoS parameters, customers are only interested in the offered services. Henceforth, customers and service providers have to negotiate agreements with respect to the provided services. These contracts are called *service level agreements* (SLAs). The SLAs, beside the technical specifications that are called *service level specifications* (SLS), contain other aspects like payment terms, durations etc…From the technical point of view, maybe the most important part of the SLS is the *traffic conditioning specification* (TCS) that incorporated the detailed service level parameters like: expected throughput, drop ratio and/or delay, traffic profile for which the requested service can be provided and the handling of out of profile traffic etc…

In order to provide end-to-end service individual DiffServ domains need to cooperate, i.e., the first domain has to negotiate with the customer while all consecutive domains must perform negotiation with each other on the behalf of the customer. Such a cooperating domain set is called DiffServ region. For an example DiffServ architecture see Figure 3.7.5. [3.7.16]

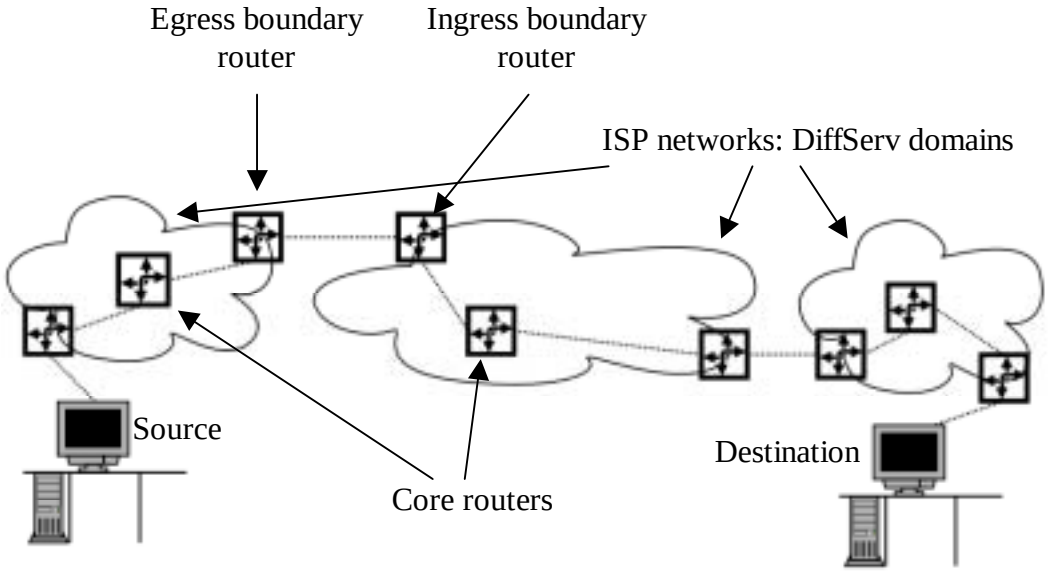


Figure 3.7.5. DiffServ Architecture

Upon the start of a service, either the end hosts or the ingress edge boundary routers mark the flows' packets in the DS field according to the agreement.

The IETF Differentiated Services Working Group defined the DS field for both IPv4 (see Figure 3.7.6) and IPv6 (see Figure 3.7.7) [3.7.2]. In the IPv4, the Type of Service (ToS) octet is used for this purpose, and in IPv6, the Traffic Class byte is defined to include the DS field. Six bits of the field are used for the code-point (DSCP) while the remaining two bits are currently unused and ignored by DiffServ-compliant nodes. [3.7.16]

<table>
<tr><td>Version</td><td>IHL</td><td>Type of Service</td><td colspan="2">Total Length</td></tr>
<tr><td colspan="3">Identification</td><td>Flags</td><td>Fragment Offset</td></tr>
<tr><td>Time To Live</td><td colspan="2">Protocol</td><td colspan="2">Header Checksum</td></tr>
<tr><td colspan="5">Source Address</td></tr>
<tr><td colspan="5">Destination Address</td></tr>
</table>

Figure 3.7.6. Header format of IPv4

<table>
<tr><td>Version</td><td>Traffic Class</td><td colspan="2">Flow Label</td></tr>
<tr><td colspan="2">Payload Length</td><td>Next Header</td><td>Hop Limit</td></tr>
<tr><td colspan="4">Source Address</td></tr>
<tr><td colspan="4">Destination Address</td></tr>
</table>

Figure 3.7.7. Header format of IPv6

The DS code point assignment also maintains backward capability with the traditional IP precedence field of IPv4 [3.7.2].

Traffic Classification and Conditioning

Once an SLS has been negotiated, the service provider has to configure traffic conditioning components at the boundary routers between any two networks including individual end-hosts too. Traffic conditioning is necessary to properly provision the network and to protect the services of overloads or unconform traffic [3.7.3]. The traffic classification policy identifies the subset of traffic, which may receive a differentiated service by being conditioned and/or mapped to one or more behavior aggregates within the DS domain. Packet classifiers select packets in a traffic stream based on the content of some portion of the packet header. Two types of classifiers are defined. The BA classifier classifies packets based on the DS code-point only. The multi-field (MF) classifier selects packets based on the value of a

combination of one or more header fields, such as source address, destination address, DS field, protocol ID, source port and destination port numbers, and other information such as incoming interface [3.7.3]. Classifiers are used to steer packets matching some specified rules for further processing. Traffic conditioning performs *metering*, *shaping*, *policing* and/or *re-marking* to ensure that the traffic entering the DS domain conforms to the traffic profile specified in the TCS (see Figure 3.7.8). Based on the conditioning traffic is classified either in-profile or out-of-profile. Out-of-profile packets may be queued until they are in-profile (shaped), discarded (policed), marked with a new code-point (re-marked), or forwarded unchanged while triggering some accounting procedure. Furthermore, out-of-profile packets may be mapped to one or more behavior aggregates that are inferior in some dimension of forwarding performance to the BA into which in-profile packets are mapped. [3.7.16]

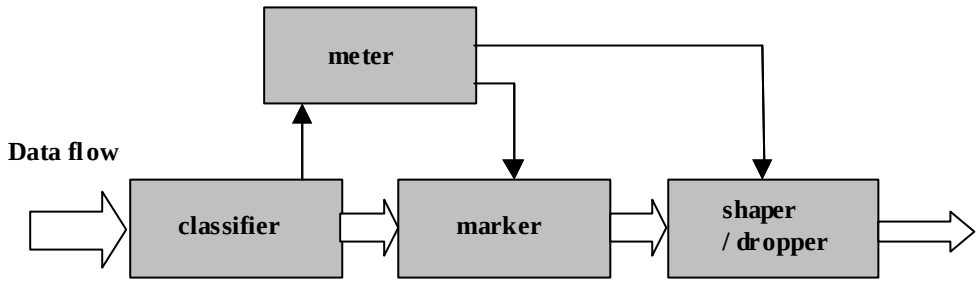


Figure 3.7.8: Traffic Conditioning in a DiffServ Node

DiffServ Scheduling and Buffer Management

Unlike the integrated services approach the diffserv framework does not specify explicitly the scheduler architecture but only the requirement of relative treatment of different traffic classes. Realizations are completely left to the hardware manufacturers. Still, there is a queueing architecture named *class based queueing* (CBQ), which seems to be a good a candidate for general diffserv scheduling. CBQ can be considered as a hierarchical, prioritized weighted round robin or WFQ scheduler (see Figure 3.7.4.) that operates on aggregated traffic classes. With CBQ, a multi-level traffic class hierarchy can be constructed where either prioritized or weight based preemption of these classes are established. Each of the class groups can be limited in accessing bandwidth or allowed to borrow from others. Figure 3.7.9 shows an example scenario for video, audio and best effort traffics like ftp and telnet.

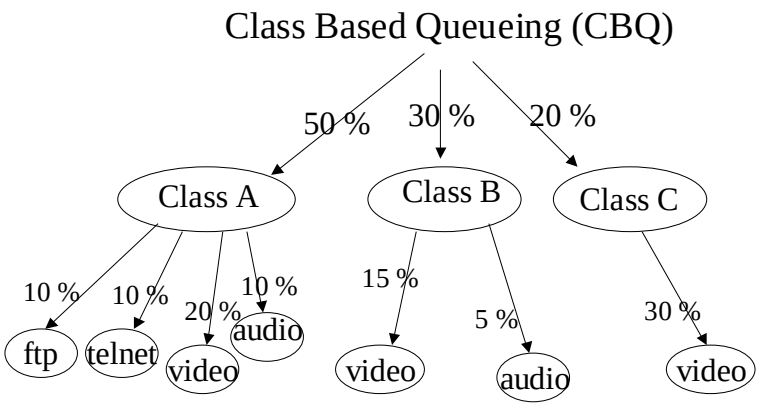


Figure 3.7.9: Class based queueing

One can notice that similar traffic types, e.g., audio, can appear at different scheduling classes whose quality certainly differ.

Categorizing DiffServ Services

In a DiffServ domain the offered services can be grouped to *quantitative* or *qualitative*. A quantitative service offers hard guarantees that can be verified by suitable measurements irrespective of any other parallel services. A qualitative service on the other hand does not provide strict guarantees but rather relative preemptions that depend on other services as well. Here, the service verification can only be done by comparison among multiple services. Another way of service differentiation is presented in [3.7.4,5], where the selection criteria are based on the methods QoS is provisioned. [3.7.16]

Service Standards

There are currently four services standardized at the IETF, namely i) default (best-effort) PHB, ii) class-selector PHB group for IP precedence compatibility, iii) assured forwarding PHB group and iv) expedited forwarding PHB. The default PHB (i) must be available in all DiffServ-compliant nodes [3.7.2] and is in compliance with [3.7.6]. Reference [3.7.2] recommends the code-point ‘000000’ for the default behavior aggregate. Moreover, any packets with unknown DSCPs should be mapped to this default PHB. The class-selector PHB group for IP precedence compatibility (ii) establishes backward capability to ToS field definitions. Henceforth, according to [3.7.2] code-points with values ‘xxx000’ should be mapped to PHBs that are best

compatible with that particular IP precedence code. These eight code points establish a relative service differentiation model. The assured forwarding PHB group (iii) has two parameters: (x) one is used to identify the precedence order while the other (y) is used to define drop preference (e.g.: AF11 and AF12 is possibly enqueued to the same buffer however AF12 will experience higher drop rates in case of congestions. Note that a realization may use different buffers for AF11 and AF12, however it must avoid packet reordering.) The available 12 AF code points are subdivided into 4 AF classes and 3 drop precedence in each class [3.7.7]. Lastly, the expedited forwarding PHB (iv) [3.7.8] provides a quantitative service with hard guarantees, i.e., low loss, low latency, low jitter and peak rate allocation up to the negotiated bit rate. This service appears as a virtual leased line to customers. According to [3.7.8], if unlimited preemption of EF traffic is allowed in the system then EF class's resource usage must be limited in order to avoid starvation of lower priority traffic. [3.7.16]

Bandwidth Brokering

Bandwidth broker (BB) entities perform dynamic management of resources in DiffServ networks (e.g. call admission control, configuration of routers, resource allocation signaling, authentication). BBs are not only responsible for the management of resources, they also control inter domain communication in order to guarantee quality of service for multi domain traffic. From this aspect, BBs can be regarded as resource agents, as they receive user requirements and decide whether they can share the available resources among the users to fulfill their requirements. If the answer is yes, they reserve the requested amount of resources for the users by configuring network nodes according to the given services. BBs are assigned to domains. [3.7.16]

Referencies

[3.7.1] Stardust Forums, Inc.: "The Need for QoS", White paper, http://www.qosforum.com/, 1999.

[3.7.2] K. Nichols, S. Blake, F. Baker and D. Black: "Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers", Internet RFC 2474, December 1998

[3.7.3] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang and W. Weiss: "Architecture for Differentiated Services", Internet RFC 2475, December 1999

[3.7.4] H. Saito, Cs. Lukovszki, I. Moldován: “Local Optimal Proportional Differentiation Scheduler for Relative Differentiated Services”, IEEE International Conference on Communication and Computer Networks, 2000, Las Vegas, Nevada, USA

[3.7.5] C. Dovrolis, P. Ramanathan: “A case for relative differentiated services and the proportional differentiation”, IEEE Network, September/October 1999, pp. 26-34

[3.7.6] F. Baker: ”Requirements for IP Version 4 Routers”,
RFC1812, June 1995

[3.7.7] J. Heinanen, F. Baker, W. Weiss and J. Wroclawski:
“Assured Forwarding PHB Group”, Internet RFC2597, June 1999

[3.7.8] V. Jacobson, K. Nichols and K. Poduri: “An Expedited Forwarding PHB”, Internet RFC2598, June 1999

[3.7.9] Weibin Zhao, David Olshefski and Henning Schulzrinne, "Internet Quality of Service: an Overview," Columbia University, New York, New York, Technical Report CUCS-003-00, Feb. 2000.

[3.7.10] S. Shenker, C. Partridge, R. Guerin, “Specification of Guaranteed Quality of Service”, RFC 2212, September 1997., ftp://ftp.isi.edu/in-notes/rfc2212.txt

[3.7.11] J. Wroclawski, “Specification of the Controlled-Load Network Element Service”, RFC 2211, September 1997., ftp://ftp.isi.edu/in-notes/rfc2211.txt

[3.7.12] S. S. Sathaye, “Traffic Management Specification Version 4.0”, The ATM Forum Technical Committee, ATM Forum/95-0013R10, February 1996

[3.7.13] R. Braden, D. Clark, S. Shenker, “Integrated Services in the Internet Architecture: an Overview”, RFC-1633, June 1994

[3.7.14] R. Braden, L. Zhang, S. Berson, S. Herzog, S. Jamin, “Resource ReSerVation Protocol (RSVP) - Version 1 Functional Specification”, RFC-2205, 1997

[3.7.15] R. Neilson, J. Wheeler, F. Reichmeyer, S. Hares, “A Discussion of Bandwidth Broker Requirements for Internet2 Qbone Deployment, Version 0.7'', Internet2 Qbone BB Advisory Council, 1999. August

[3.7.16] Császár András, “Differentiated Services for Voice Communication”, Master’s Thesis, Budapest University of Technology and Economics, 2001

3.8. Switching in mobile networks

Gusztáv Adamis, author

Katalin Tarnay, reviewer

The main difference between the switching used in mobile and in fixed networks is the handling of the mobility of the subscribers. For this purpose special equipments and special protocols are used. The most important equipments in connection with the management of the mobility are the HLR (Home Location Register) and the VLR (Visitor Location Register). There is one HLR in every mobile network, containing all pieces of data about each subscriber of that network, including its location. To each mobile switch (MSC) a VLR is connected to, containing the information about those subscribers who are located at the area served by the MSC. When a subscriber performs a Location Update in a given MSC-area, the MSC informs the HLR of the subscriber about it, and assigns a temporary telephone number (MSRN - Mobile Station Routing Number) to the MS. The management of the mobility between the MSCs and other elements of the network (e.g. HLR, VLR, etc.) is controlled by the MAP protocol, that is part of the CCSS7 architecture and uses its lower level protocols. But the establishment of the call itself is managed by the ISUP protocol.

Between the MSC and the mobile station several functional units, several transmission media and several protocols are used.

For the control between the MSC and BSC (A interface) the BSSMAP protocol is used, whose main purpose is the assignment of radio channels to the calls. The A interface is also used for transparent transmission of messages between the MSC and MS used for Call Control (CC) and Mobility Management (MM), constituting the DTAP protocol.

The Abis interface is located between the BSC and BTS. The signals here are carried by the LAPD protocol - similarly to the DSS1 subscriber loop protocol of the ISDN - and not by the CCSS7, while on the Radio interface its modified version, the LAPDm is used, that is optimised to radio transmission (main difference is the much

shorter length of the messages). Here, on the subscriber side, for the management of the radio resources, the RR protocol is used.

In next chapters we give an overview about the most important protocols, and finally we give an example how they are used.

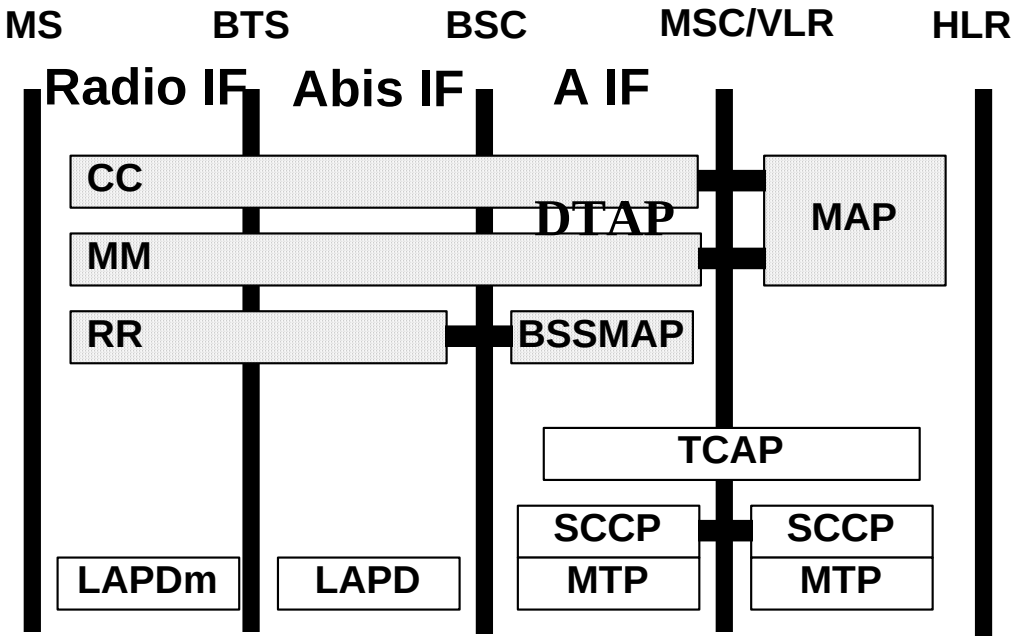


Figure 3.8.1.The structure of the signaling system no. 7 for radio connections

SCCP

While the MTP is suitable to transmit signalling traffic between any two signalling points only of the same signalling network, the SCCP (Signalling Connection Control Part) is suitable to perform signalling between any two signalling points even of different signalling networks. What makes it possible? The SCCP has the routing capability based not only on the CCSS7 point codes, but it can also use the Global Titles for routing. Global Titles are identifiers that are out of scope of the CCSS7 (like telephone numbers, IMSI, etc.). Because the Point Codes are 14 bits long, they allow to address some 15 thousands of signalling points world-wide, that is too few. So every service provider routes its internal signalling traffic on its own signalling point code scheme, and only some of its signalling points are connected to the international signalling network. These gateway signalling points have two signalling point codes: one of them is used in the international signalling network,

while the other is used in the local signalling system of the given operator/service provider. The MTP SIO octet determines which addressing scheme to be used. If we need a signalling connection where the signalling points involved belong to the same signalling network, then the SCCP routes on the point codes. But, if the signalling points are in different networks (e.g. if we want to reach the HLR of an other operator), then the SCCP would use the telephone number of the HLR in this example, as a global title. First it determines - on the basis of the telephone number - the proper gateway SCCP of the local signalling network toward the target network. The SCCP asks the routing to that gateway signalling point from the MTP, based on the local point codes. The Gateway SCCP determines - still on the basis of the global title - which is the incoming gateway signalling point of the target network, and routes the message toward it by the international point codes, while the Gateway SCCP of the target network decides - still by analysing the global title - where to route the message within the local network using the local point codes.

The SCCP is also suitable to make differences between several subsystems, like HLR, VLR, MSC, etc. The SCCP provides two connection-oriented and two connectionless services, though the GSM uses only one of the latter.

The connectionless service is used for the transaction-oriented services (e.g. for asking the HLR for a piece of data). In this case the SCCP is used only for its addressing capability and for providing an interface to reach the signalling network, for managing the transactions (e.g. to find the answer for a request) is the task of an other protocol, of the TCAP.

The connection-oriented services of the SCCP are used by the BSSAP protocol, when a signalling connection should be established e.g. between a mobile station and an MSC in Location Update. The steps of the connection establishment are very similar to those of the ISUP, but the messages involved are different (CR - Connection Request, CC - Connection Confirm, DT1 - Data Form 1, RLSD - Released, RLC - Release Complete), and the assignment of the signal and the (voice) connection is based on logical identifiers, the randomly assigned SLR/DLR (Source/Destination Local Reference) instead of physical ones, the circuit identifier codes.

TCAP

For controlling the GSM calls, transactions are used. For the transactions there is no need to establish signalling connections, but the transactions must somehow be identified, and if a transaction has a request/answer form (as in most of the cases), then somehow it must be able to pair the proper messages. Among others, these are the tasks of the TCAP (Transaction Capabilities Application Part) protocol, that uses the UDT (UnitData) message of the connectionless service of the SCCP. The SCCP is used here only as an interface to the CCSS7 (SCCP addressing, interface to MTP), it does not give any further help to the management of the transactions, this is the task of the TCAP on its own.

The TCAP can handle dialogues (transactions) that consist of one or several operation execution. In according to it a TCAP message may consist of two main parts: one dialogue and one or several component portions.

The following message types are defined in TCAP: BEGIN (to start a dialogue), END (to close a dialogue in a normal way), CONTINUE, and ABORT. A dialogue typically consists of a BEGIN/END pair, but if the request and/or the answer cannot fit in one message, or if more operations should be done in a transaction, any side may include the necessary number of CONTINUE messages. The abnormal termination of a dialogue may be indicated by ABORT messages. If the source of the error is the TCAP or the lower levels, a P-ABORT (provider), or if it is the user (typically the MAP) of the TCAP, a U-ABORT message is used. The message type code is followed by the transaction identifiers (OTID/DTID - Originating/Destination Transaction Identifier), whose purpose are similar in identification of the transactions to the role of the SLR/DLR in the SCCP in the identification of the connections.

The message type and the transaction identification may be followed by a dialogue portion. It contains the identification of the TCAP version, the identification of the protocol and its version that uses the TCAP, and it is possible to convey any user - that is non-CCSS7 related - information. If we do not want to send this, or if it is obvious which TCAP and TCAP user protocol is used on a given signalling link, this part may be omitted.

This part is followed by one or more Component Portions, that carry the messages of the TCAP user protocols (in the case of GSM it can be MAP or INAP).

These components may be INVOKE, RETURN RESULT LAST/NOT LAST, RETURN ERROR or REJECT. An INVOKE component requests an operation to be performed, whose result is carried by a RETURN RESULT LAST, or if it is not enough, it is preceded by the appropriate number of RETURN RESULT NOT LAST components. A RETURN ERROR component is sent, when the request was understandable, but it could not been able to be performed for any reason (e.g. the requested data was missing from the data base), while the REJECT component indicates that the request was not understood.

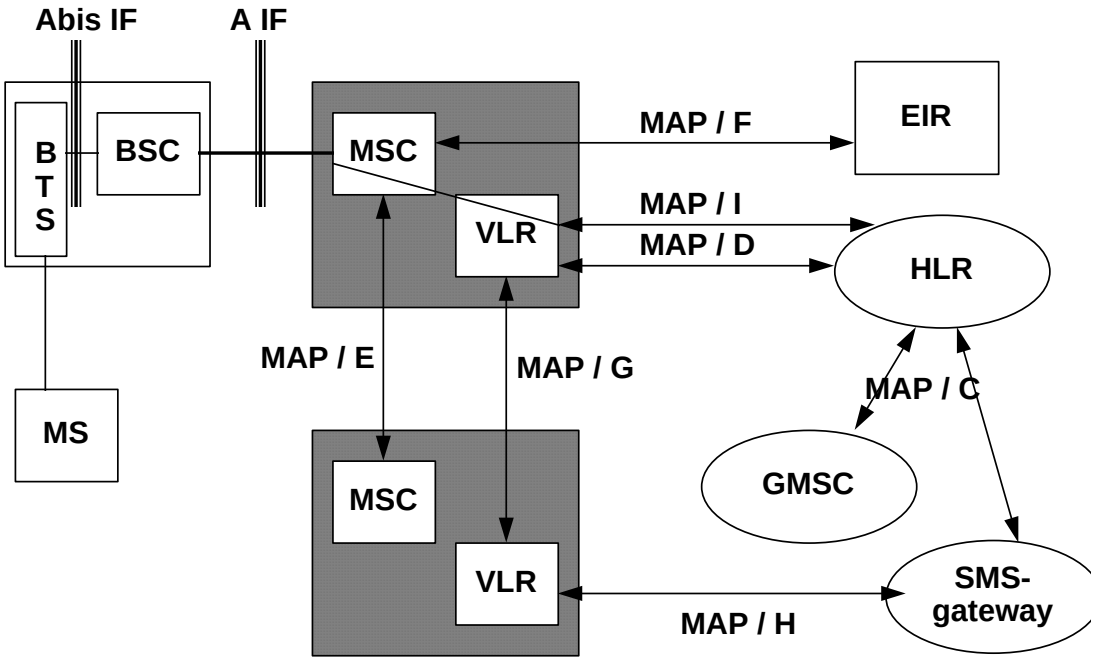


Figure 3.8.2. Signaling transmission in the GSM network solving routing tasks

MAP

The MAP (Mobile Application Part) protocol is used for gathering the appropriate pieces of information to establish mobile calls and for the management of the mobility. The MAP uses transactions, so it is a typical user of the TCAP.

The MAP can be functionally divided into several subprotocols. The B interface describes the connection of an MSC and a VLR, but since they are always implemented as one functional unit, it is not implemented in practice.

The C interface specifies the messages between an HLR and Gateway MSCs or SMS Gateways. These messages are used to ask the HLR for the routing information (MSRN).

The D interface contains the messages between the VLR and the HLR. These messages are mainly used to inform the HLR if a mobile station performs a Location Update in a new VLR-area, and to inform the old VLR about it by the HLR, to clear the data of the subscriber The messages that are used by the HLR to ask the actual VLR for the routing information (MSRN, that will be transferred to the Gateway via the C interface) are defined here, too. The other task of the D interface is the management of the Supplementary Services (switching them on/off, sending passwords, etc.). The related messages are grouped on the Figure into the - not standardised - I interface.

The E interface specifies the messages between MSCs. These are mainly in connection to the handling of inter-MSC handovers, and for sending SMS-es. The latter is indicated as the - not standardised - H interface.

The only operation of the F interface is the IMEI checking between the MSC and the EIR, while the only operation of the G interface is used by an MSC to get the authentication keys from the old MSC when a mobile station changes MSC-area.

The MAP is used to follow the location of a mobile station and to gather the routing information (MSRN), but the call itself will be established by the ISUP in a similar way as in fixed networks, but the MSRN will be used instead of the dialled number.

We have to mention here, that the messages of the MAP and the TCAP are encoded by the BER (Basic Encoding Rules) defined in ASN.1, and even for the specification of the MAP messages the ASN.1 (OPERATION and ERROR) macro feature is used.

The protocols of the A interface

The A interface is located between the MSCs (that perform the role of the telephone switches in GSM networks) and the BSCs (that control the transceiver stations). The protocol of the A interface is called as BSSAP (Base Station Subsystem Application Part), that is the unification of the BSSMAP (Base Station Subsystem Management Application Part) and the DTAP (Direct Transfer Application Part) protocols. The BSSMAP is used for controlling the radio resources of the BSCs, while the DTAP is used for managing the mobile stations and the mobile calls, its role

- apart from the mobility-specific requirements - is similar to the role of the DSS1, the subscriber loop protocol of the ISDN.

The protocols of the BSSAP use the MTP and the SCCP protocols of the CCSS7. The DTAP uses mainly the connection-oriented services of the SCCP - since its main purpose is to establish calls -, while the BSSMAP uses mainly the connectionless services. Difference between the messages of the two protocols can be made by a discriminating octet at the beginning of the data field of the SCCP messages.

The connectionless messages of the BSSMAP are similar to the Circuit Supervision messages of the ISUP, but they are used for managing radio channels. An other typical connectionless BSSMAP message is the Paging message, that is used in Mobile Terminating (MT) calls. The call termination is completely different than it was in the fixed networks. Here the called party gets an indication (this is the Paging) about a call to be terminated to it. After receiving it, the mobile station itself establishes the connection, similarly when it really wants to originate a call (of course the reason of the call establishment is indicated in the proper message).

An other important task of the BSSMAP is the management of the handovers.

BSSAP uses connection-oriented messages during the connection establishment, too. The most important ones are: the first message of the call establishment, the radio channel assignment to the call and the ciphering messages.

The DTAP messages are used to establish and release calls between the mobile stations and MSCs. The DTAP can be functionally divided into three parts: the first part is the Call Control, that is functionally similar to the DSS1. The Alerting, Setup, Call Confirmed messages are used for call establishment, while the Disconnect, Release, Release Complete for call clearing. The other part of the DTAP is the Mobility Management. Its main tasks are: initiation of Location Update, authentication of mobile stations, TMSI assignment and modification, and the indication of switching the mobile stations off (IMSI detach). The third part contains the SMS-related messages.

Establishment of mobile calls

Finally, after looking through the protocols of the GSM let us see how a mobile call is established. In the example the steps of a PSTN to mobile call can be followed.

The first step is the arrival of an ISUP IAM message to the proper gateway of the mobile network, containing the telephone number (MSISDN) of the subscriber. The task of the gateway is to route the call to the MSC where the subscriber is currently located. But, to do that, it must collect the temporary number (MSRN), which the MS can be accessed on. This number was assigned by the MSC when the MS performed a Location Update to it. So the GMSC asks the HLR, that knows in which MSC-area the MS is currently located, so the HLR is able to collect the MSRN from the MSC (or actually from its VLR), and transmits it to the GMSC. This way the GMSC is able to establish the call, using the MSRN on ISUP level.

When recognising the call attempt, the MSC sends a Paging message to the MS, asking it to make a call. It is responded by a Paging Response embedded in a Complete Layer 3 Info message. (This is how the MSC can know the reason of the call initiated by the MS.) Authentication follows, and they switch to ciphering mode.

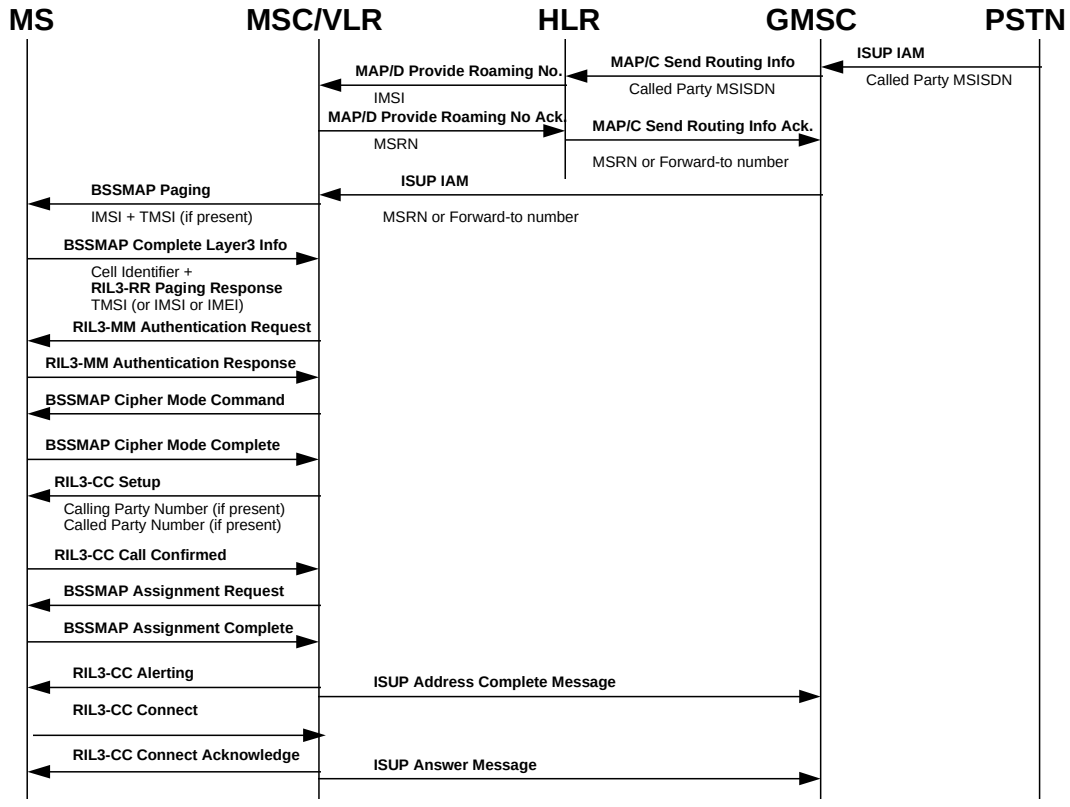


Figure 3.8.3 ISUP type signalling process

Finally, the call will be established on the subscriber side in a similar way, as it is in the DSS1. The main difference is the Radio Channel Assignment. When the call is answered (Connect), the MSC sends an ISUP ANM message to the GMSC, that will transmit it to the switch of the calling party.

References

[3.8.1] Mouly - Pautet: The GSM System for Mobile Communications ISBN 2-9507190-0-7 1992.

[3.8.2] ETSI GSM 09.02 Mobile Application Part Specification

[3.8.3] ITU-T Recommendation Q.771 - Q 774 TCAP

[3.8.4] ETSI GSM 04.08 Mobile Radio Interface Layer 3 Specification

3.9. Development of telecommunications protocols

Gusztáv Adamis, author

Péter Kesselyák, reviewer

The protocol engineering, a new branch of science deals with the questions of the production of telecommunication protocols. The main steps of the development of a protocol are the followings:

First a verbal (nonformal) specification is developed. Because of the usage of a human language the specification may contain errors, may not be complete, may be ambiguous, and in order to be able to support the further steps of the development process the specification must be given in a formal way by using some Formal Description Technique (FDT - is a specification language with formal semantics).

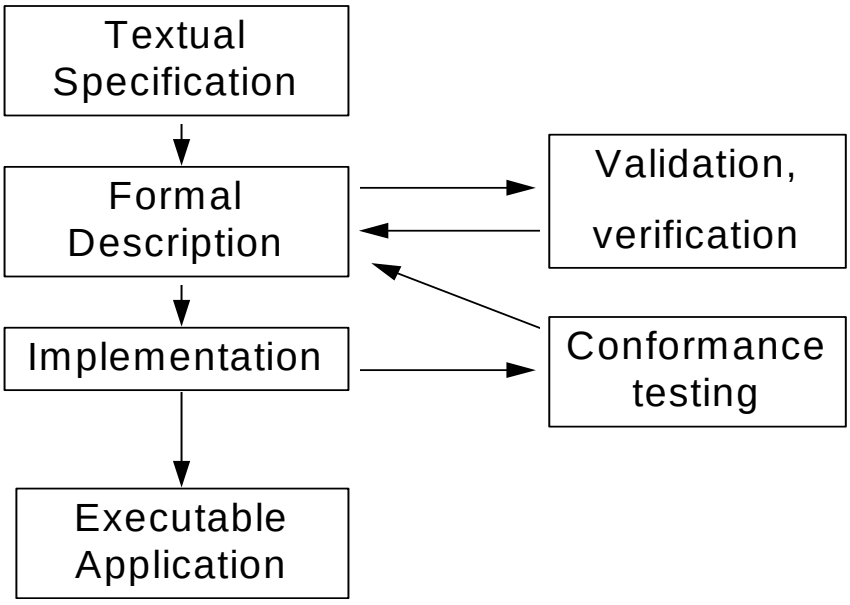


Figure 3.9.1 Steps of protocol engineering

Next step is the validation & verification - that is to decide if the specification is correct (no deadlocks, unreachable or trap states etc.) and contains all the same as the informal specification. Of course, according to the detected errors the specification must be modified.

The next step is the implementation of the specification. In the implementation or during the specification the simulation of the protocol often helps.

Finally it must be tested if the implementation implements really the specification. This step is called as Conformance Testing - it checks wheter the implementation conforms to the specification or to a standard. This step is often performed by independent, authorised test laboratories. (The details of the testing can be found in Chapter 3.10; this chapter deals only with the specification and implementation.) Of course, according to the detected errors the implementation must be modified.

Specification of protocols

The specification of a protocol has high importance, since the further steps of the development are based on it. This is why the quality of the implemented protocol and the cost of the development process highly depends on the way of the specification. The specification must be:

(1) Unambiguous, that is every implementor must understand it in the same way;

(2) Complete, that is there must be instructions for every possible situation, not only for the desired reactions but for the handling of the erroneous situations, too.

(3) Easy to validate, possibly in an automatical way supported by computers;

(4) Implementable, possibly in an easy, straightforward, automatical way.

These requirements can be fulfilled only by the formal description techniques. The formal description techniques have well-defined syntax and formally defined semantics.

The most commonly used model of a protocol is based on the communicating finite state automata. Each process of a protocol is represented by a finite state automaton, and they are communicating with each other by messages via communicating channels. The channels are represented by errorless FIFO queues.

But for real protocols this model is often not suitable, since the number of the states is great. The extended finite state automaton (EFSM) helps on this problem, that can be represented like this:

EFSM = <S, I, O, V, P, A, s0, f(s,i,p)>, where

S: finite set of states

I, O: finite set of input/output messages

V: finite set of the variables

P: finite set of the predicates (conditions)

A: finite set of actions (e.g. assignment a value to a variable)

$s0 \in S$: start state

f(s,i,p): state transition function that describes if the EFSM is in $s \in S$ state, receives an $i \in I$ message when a $p \in P$ predicate is true, then which $a \in A$ action is to be done, which $o \in O$ output message is to be generated and which $s' \in S$ state will be the next state.

This model is suitable to describe real telecommunications protocols. The model is quite lifelike, that's why it is easy to understand and use. A lot of FDTs are based on the communicating EFSM model. Among them the SDL (Specification and Description Language) developed by CCITT/ITU-T has outstanding importance in the field of telecommunications.

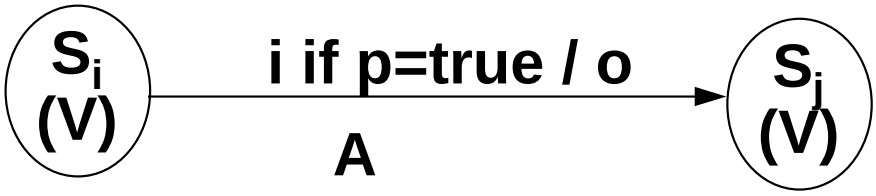


Figure 3.9.2. Extended finite state automaton

The SDL language

The SDL is a language based on the Extended Finite State Machine concept. In the SDL the *system* that communicates with the environment, can be divided into *block*s, that are interconnected by delaying (or possibly nondelaying) *channel*s. The blocks can be further refined step-by-step, while we finally reach the *process*es, that are actually communicating EFSMs. The processes within the same block, as well as the processes and the channels are interconnected by nondelaying *signalroute*s. Via the signalroutes and via the channels the information is carried by *signal*s. There belongs an infinite incoming FIFO queue to each process, where the incoming signals are gathered till they are processed. The SDL allows to define *timer*s. The data definition part of the SDL is based on the abstract data types (ADTs). In the newer versions - among many other extensions - a lot of object-oriented features were introduced.

The SDL is the only one among the standardised formal description languages, that has a graphical version - or actually it had a graphical version (SDL/GR) first, and only when it had a great success was the programming language like (SDL/PR) version developed.

The philosophy of the SDL - especially of its graphical version -is close to the practical life. The main advantage of the language that it makes it possible to define a system in a hierarhical way: from the top level overview to the implementation close description.

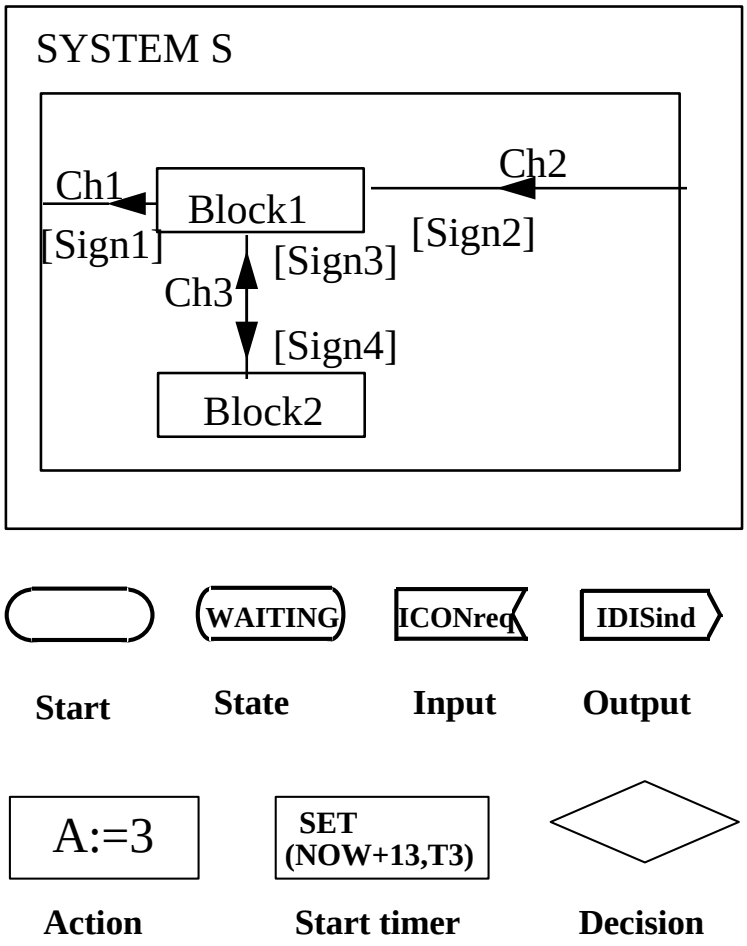


Figure 3.9.3. Some important SDL symbols

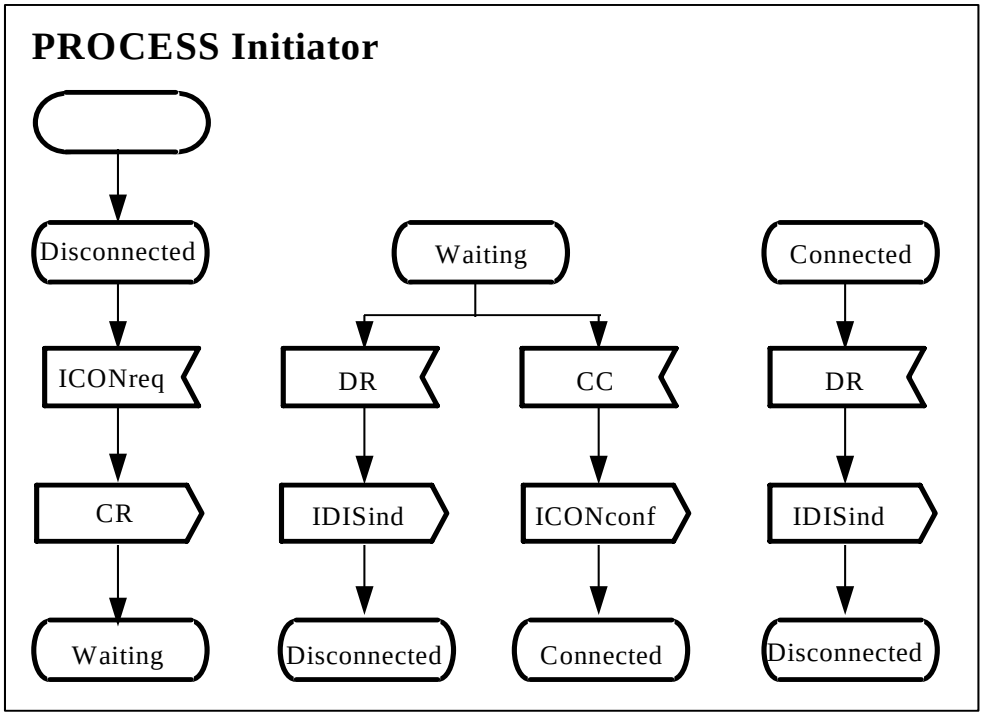


Figure 3.9.4. Description of an SDL process (illustration)

Specification of data structures

SDL uses an abstract data type description method for data type specification. This method is highly suitable to precisely define data structures in an implementation independent way. This is the advantage, but also the disadvantage of this method, since it makes difficult the automatical implementation and requires a way of thinking different from that of most other languages.

This is why there is an other way to define data structures in SDL: by the help of the ASN.1 (Abstract Syntax Notation One) language. This language is close to the data specification features of high level programming languages. An other advantage of the ASN.1 is that it can also be used in the TTCN test specification language, so there is no need to define data in two different ways at test specification.

The ASN.1 was originally developed for specifying the unambiguous and implementation-independent data communication of application layer protocols, where the protocol automaton is typically fairly simple (with a few states only), but the structure of the messages is rather complicated, they typically contain a lot of information elements of complicated data types, that may be optional or may have default values.

When specifying the data communication we have to provide the followings:

- unambiguous specification of (possibly very complex) data structures
- unambiguous encoding/decoding of the messages that makes it possible at the receiving side to decide which information elements of the messages were actually transmitted and with which values, even in the most complicated situations.

So the ASN.1 also has two main parts: a data specification and an encoding rule specification one.

The data specification part of the ASN.1 is similar to that of most high level programming languages. On one hand it defines several built-in types, and on the other hand it makes it possible to create more complicated types from simpler ones. These constructional features though are more flexible to those of typical programming languages. But the ASN.1 is not a programming language, so it does not include “executable” instructions, like cycles, if-sentences, etc. ASN.1 does not specify the implementational details (like size, range, etc.) of its built-in types. These questions are left for the encoding.

We can create new simple types from the existing ones by renaming, and for the creation of structured types the following construction rules are available: SEQUENCE, SET, SEQUENCE OF, SET OF, CHOICE.

A SEQUENCE/SET produces a record/set of elements of possibly different types (they are similar to the records or structures of high level languages), while SEQUENCE OF/SET OF defines a record/set of elements of the same types (they are close to the arrays of other languages). The transmission order of the elements of SEQUENCEs must be the same as it is given in the specification, while for SETs it is not defined. By CHOICE a C-union like structure can be created, where only one of the specified fields is presented at a given time.

Subtypes can be generated from the previously defined types by restrictions. Some examples:

```
TelephoneNumber :=
                    IA5String(FROM('1'|'2'|'3'|'4'|'5'|'6'|'7'|'8'|'9'|'0'|'*'|'#'))

NumbersLessThanTen:= INTEGER(0..9)

ArrayWith1to64Elements:= SEQUENCE (SIZE (1..64)) OF INTEGER
```

ArrayWithExactly64Elements:=SEQUENCE (SIZE (64..64)) OF INTEGER

SmallPrimes:= (2|3|5|7)

Sometimes specifying the types of the data elements is not enough to decode them unambiguously. Think about a record with two fields of the same data type, that are optional, and in a given situation only one of them should be transmitted. To be able to identify at the receiving side which of them was sent, the ASN.1 gives the possibility to assign unique data type identifiers to the field, these are called as TAGs. This is how we can create different type fields from the same type.

Coordinates:= SEQUENCE{

x [0] INTEGER OPTIONAL,

y [1] INTEGER OPTIONAL}

In the example above we have two optional INTEGER fields, but the first is a [0]-kind-of-INTEGER, and the other is a [1]-kind-of-INTEGER, where the numbers in [ ] are the TAGs.

ASN.1 defines unambiguous encoding rules. The essence of them is to send - at every data transmission - the type, the length and the value of the given data element.

Implementation of protocols

One of the key questions of the protocol engineering is the implementation, since the goal of the development is to bring the specified, validated protocol to a working software or hardware form, that satisfies all the requirements defined in the specification.

The implementation of a protocol starts by the specification. The specification - though it has great importance - on its own is not enough for implementing a protocol, since the protocol specification should be more or less implementation-independent. (Because we should be able to implement the same protocol on different platforms and/or operating systems.) So during the implementation process we have to make decisions that are not explicitly given in the specification.

When implementing a protocol the implemented functions can be divided into two more or less disjunct subsets. The first subset contains the part of the protocol

functions that can be derived directly from the specification. But some other parts will be determined by the features of the target machine or the operating system. This second subset contains routines that provide the connection between the protocol and the target environment, like: memory management, queue management, real-time support, control of the communication among tasks, scheduling of processes, event processing, error handling and recovery routines. This part is mainly determined by the features of the target environment and not by the protocol itself. Of course there must be an interface between these two subsets.

There are two main ways of implementing protocols.

Traditionally, the protocols are implemented in a 'manual' way - that is a group of programmers develops the implementation from the specification as an independent software. The main advantage of this method, that since the developers

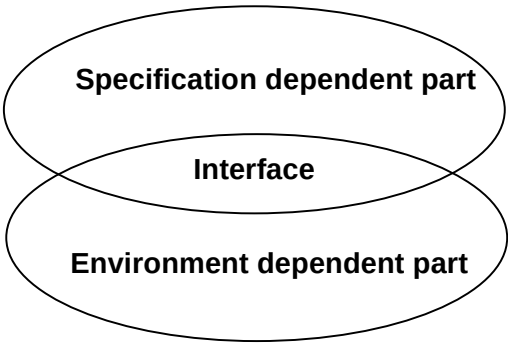


Figure 3.9.5. Classification of the routines of a protocol implementation

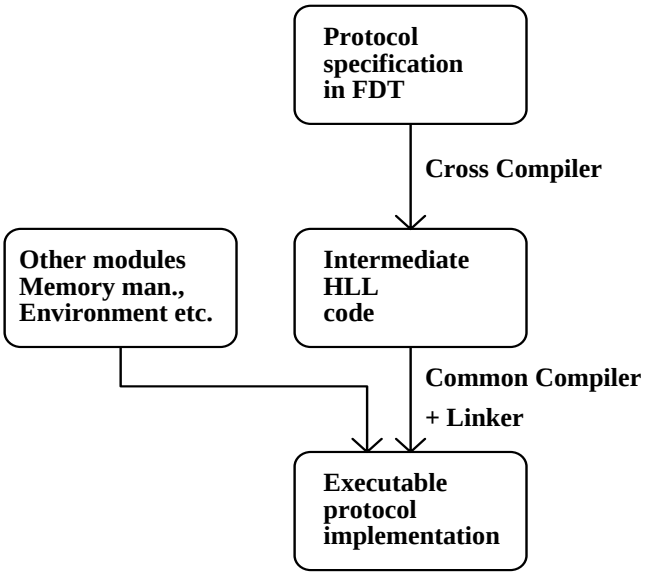


Figure 3.9.6. The process of the automatical protocol implementation

can take into consideration both the requirements of the protocol and the features of the target environment, the code can be optimised to the specific protocol and the target environment, so it may be fast and may have a (relatively) small size. The main disadvantage of this method is that the development is quite time consuming, especially the testing, the finished program is less robust, hard to maintain and typically not portable.

The other approach is the (semi) automatic development. It means that (the major part) of the implementation is automatically generated from the specification by a cross compiler - if the specification was created in an FDT. But because apart from the specification-dependent parts there are environment-dependent parts as well, this automatic code generation cannot be complete, some - typically manually developed - environment related modules must be added to it. To make it easier, this process is most commonly carried out in two steps as can be seen on Figure 6.

The comparison of the methods is in Table I.

	AD-HOC	SEMI AUTOMATICAL
	protocol implementation	
Running speed:	high	lower
Size of code:	small	larger (2x or more)
Structure of code:	complicated	structured, simple
Modularity:	typically small	high
Portability:	low	high
Maint., further devel.	complicated	simple
Debugging:	complicated, time consuming	simple
Development time:	long	short
Conformance to the specification	must be checked	automatic

Table I. Comparison of the protocols implemented in 'ad-hoc' and semi automatical ways

References

[3.9.1] ITU-T Recommendation Z.100: Formal description techniques (FDT) - Specification and Description Language (SDL), 1999.

[3.9.2] ITU-T Recommendation Z.105: SDL combined with ASN.1 (SDL/ASN.1), 1994.

[3.9.3] ITU-T Recommendation X.680 Abstract Syntax Notation One, 1994.

[3.9.4] ETSI EG 202 106 (1999): Methods for Testing and Specification (MTS): „Guidelines for the use of formal SDL as a descriptive tool".

[3.9.5] ETSI EG 202 107 (1999): Methods for Testing and Specification (MTS): „Planning for validation and testing in the standards-making process".

[3.9.6] ETSI TR 101 680 (1999): Methods for Testing and Specification (MTS): „A harmonized integration of ASN.1, TTCN and SDL".

3.10. Testing Protocols and Softwares

Sarolta Dibuz, author

Katalin Tarnay, reviewer

3.10.1. Basics of software testing

More than fifty percent of the development cost of telecommunication software is the cost of testing. The aim of the testing process is to reveal as many errors of the software as possible. We can never prove that a software is errorfree. Our aim is to give the software inputs and detect if the outputs, or answers of the software is correct. The correct output can be determined from the specification of the software or from the user expectations. Testing requirements vary substantially based on implementation technologies used, and development practices.

Testing process consists of three part:

- Test planning
- Test specification
- Test execution

A Test Plan is a detailed project plan for testing, covering the scope of testing, the methodology to be used, the tasks to be performed, resources, schedules, risks, and dependencies.

A test specification defines exactly what tests will be performed and what their scope and objectives will be. A test specification is produced as the first step in implementing a test plan, prior to manual testing or automated test suite development. It provides a repeatable, comprehensive definition of a testing campaign.

Test execution comes after development of an automated test suite. Testing can address black box testing of installation, compatibility, functionality, and reliability, as well as white box testing.

There are automated test development tools that are used for implementation of a test suite following test planning and test specification. Several types of testing is possible on a software depending on the software capabilities we would like to test:

Functional testing: Validating an application conforms to its specifications and correctly performs all its required functions. This entails a series of tests which perform a feature by feature validation of behavior, using a wide range of normal and erroneous input data. This can involve testing of the product's user interface, database management, security, installation, networking, functional testing can be performed on an automated or manual basis.

Performance testing: Performance testing can be applied to understand the software's scalability, or to benchmark its performance in the environment of third party products such as servers and middleware. This sort of testing is particularly useful to identify performance bottleneck. Performance testing generally involves an automated test suite as this allows easy simulation of a variety of normal, peak, and exceptional load conditions. Manually it is usually very copmlicated to achieve big load situations in a test environment. It is necessary to define performance targets and metrics for identifying performance bottlenecks during performance testing.

Regression testing: Similar in scope to a functional test, a regression test allows a consistent, repeatable validation of each new release of a product. Such testing ensures that reported product defects have been corrected for each new release and that no new quality problems were introduced in the maintenance process. Though regression testing can be performed manually an automated test suite is often used to reduce the time and resources needed to perform the required testing.

Conformance testing: Verifying implementation conformance to industry standards. Producing tests for the behavior of an implementation to be sure it provides the portability, interoperability and compatibility a standard defines.

Interoperability testing: Means testing the interoperability capabilities of two implementations of the same specification. It can reveal design inconsistencies in the standard as well as compatibility problems of the implementations. Conformance testing is executed usually before interoperability testing to reduce the number of interoperability problems.

3.10.2. Testing protocols

Protocols are special types of software they implement the communication part of telecommunication softwares. Protocols define the communication of two nodes or elements of a network, there are several protocols each of them is designed for certain functionalities in the network. The protocols defines the:

- the syntax of the messages changed during communication,
- the semantics of the communication, what message send as an answer to any incoming message,
- and the timing behaviour of the communication i.e. the timeout values.

We saw in section 3.9 that a protocol has an engineering life-cycle and conformance testing plays important role in it. The same protocol models can be used as a base of protocol testing that are used for protocol specification. The main testing types that can be applied to protocol implementations are:

- functional testing.
- conformance testing,
- interoperability testing, and
- performance testing.

Protocol testing methods are usually black-box approaches which mean that the implementation is considered as a black box, we send messages to the black box during testing and examine the answeres received as a reply to the sent messages. We cannot see the states and the internal structure of the implementation.

3.10.3. Conformance testing and test standardization

Protocols are defined in standards. There are also standards defining test methods and test suite that can be executed during testing. Test standardization mainly concentrate on conformance testing. Testing can never be done for 100%, we have to make always a trade-off between test coverage, to what extent we test and test expenses. Test standardization defines methods for testing if an implementation fulfills requirements on the protocol interfaces defined in the protocol standards. Standardizing test requirements, test procedures and test suites make it possible to produce repetable and comparable test results as all vendors use the same test

procedure and test suite to test its implementation. This is very important when interoperability capabilities of different vendors' products has to be ensured.

ISO standardized the conformance testing framework in the standard ISO 9646. ITU-T took over this standard in the X.290-296 series of standards. These standards define:

- Test architectures, so called abstract test methods of conformance testing
- Test documents produced for and during conformance testing
- TTCN (Tree and Tabular Combined Notation) a test notation for describing test suites
- Requirements on the means of testing, the test equipments with which the tests are executed, and
- The test execution process called the conformance test campaign.

Most of the test suites and test documents are standardized by ETSI (European Telecommunication Standardization Institute). The test language is standardized so as it can be executed with different test equipment. ETSI produces standardized test suites in TTCN besides protocol standards.

Conformance testing means testing if a protocol implementation conforms to the standard. The protocol standard defines the features and capabilities of the protocol, these can be identified as conformance requiremets. During conformance testing we test if the protocol meets these conformance requirements. As all vendors' implementations have to be conform to the protocol standard, conformance testing increases the probability that different implementations are able to interwork.

Not only the conformance testing methodology is standardized but there are also standardized test suites for the protocols. Why is it so important to standardize conformance testing? Because this makes it possible for different vendors to follow the same test procedure with the same test suite. So even if the test equipment is usually different (not to tie too much the hands of the vendors), and different laboratories execute the conformance test the test results are really comparable for the implementations of the same protocol.

Prior to the test suite writing test purposes can be identified based on the protocol standard. These are textual description of well-defined objectives of testing, each test purpose focuses on one or a set of related conformance requirements. ETSI collects test purposes for a given protocol in the Test Purpose and Test Suite

Structure standard document (TSS&TP). It is also possible to describe test purposes with MSC (Message Sequence Chart).

Test cases are derived based on test purposes, one test case for each test purpose. So the test case is the implementation of a test purpose in a test language. The standard of conformance testing contains the definition of a test language called TTCN (Tree and Tabular Combined Notation). The derivation can be done automatically or manually. The automatic test case derivation is based on test generation methods. All test purposes implemented in test cases comprise the Abstract Test Suite (ATS). The ATS is standardized usually in order to get comparable and repetable test results so called verdicts. All vendors can use the same abstract test suite of the protocol so if their protocol implementation passes all test cases (fulfills all conformance requirements) this means that it conforms to the standard. The test suite is called abstract because it is independent of the protocol implementation so as all vendors can use it. This implies that the ATS is not executable without further parameterization and compilation. Ideally the ATS should be complete, that is the implementation passes it if an only if the implementation is correct. It is usually not possible to construct a complete test suite with finite number of test cases, so an ATS should be sound meaning that all implementations that do not pass the ATS are not correct. A sound test suite has to be exhaustive, so all passing implementations conform to the protocol standard.

There are other documents in conformance testing that help to produce comparable and repetable test results of different implementations, these are the PICS, PIXIT, PCTR proformas. The proformas filled in with data before and during confoemance testing are called PICS, PIXIT and PCTR accordingly.

The PICS (Protocol Implementation Conformance Statement) is a statement of the capabilities and parameter values realized in a particular protocol implementation (i.e. timer expiry values). The protocol standards produced in ETSI, ITU-T usually contain a PICS proforma that is an empty PICS. The proforma is a list of capabilities in the form of a table. For all capabilities a reference is given to the standard where it is described in more details, and there is a classification stating if it is mandatory, optional or conditional to support the capability in an implementation. The implementor fills in the PICS proforma with ‘yes’ or ‘no’ stating if the given capability has been implemented or not. During conformance testing the test laboratory selects

test cases from the ATS for execution based on the PICS as it shows which protocol features are actually implemented.

The parameterization of an ATS is done in the test laboratory based on the PIXIT document (Protocol Implementation eXtra Information for Testing), as it contains the values of those parameters that depend on the implementation. This document also describes the circumstances of testing to be able to repeat it with producing the same test results.

The test results of a protcol implementation are collected in the document called PCTR (Protocol Conformance Test Report). It lists all test cases in an ATS and indicates which of them were executed in the test campaign and gives the results as the final verdict (Pass, Fail, Inconclusive). If more protocols have been tested in a system a summary document called SCTR (System Conformance Test Reort) is also produced. Production of these documents ends the conformance testing process that is done in a test laboratory. The certificate about conformance is produced by an authority and not by the test laboratory, on the other hand of course it is based on the test results produced by the test laboratory. The authority trusts the test results produced by the laboratory as it regularly checks if the laboratory executes conformance testing in the required way.

The execution of the test suite in the test laboratory on an actual implementation is called test campaign. It has the following steps:

- Preparation for testing,
- Test operations, and
- Test report production.

The preparation for testing involves the production of the PICS and PIXIT documents. PICS is filled in by the implementor and the PIXIT is filled in by the test laboratory, with some additional data given by the implementor. The operations part involve the static conformance review, which means the analysis of the PICS. The review checks if all mandatory capabilities have been indicated as implemented and the answers for all the optional and conditional capabilities are consistent. Then comes the execution of the ATS. This is called dynamic testing because in this phase the real message exchange is taking place between the test equipment and the Implementation Under Test (IUT) (Figure3.10.1). The test report production process

provides the clients of the laboratory with documentation they can use for further steps such as obtaining a certificate from the authority.

3.10.4. Languages for describing test suites

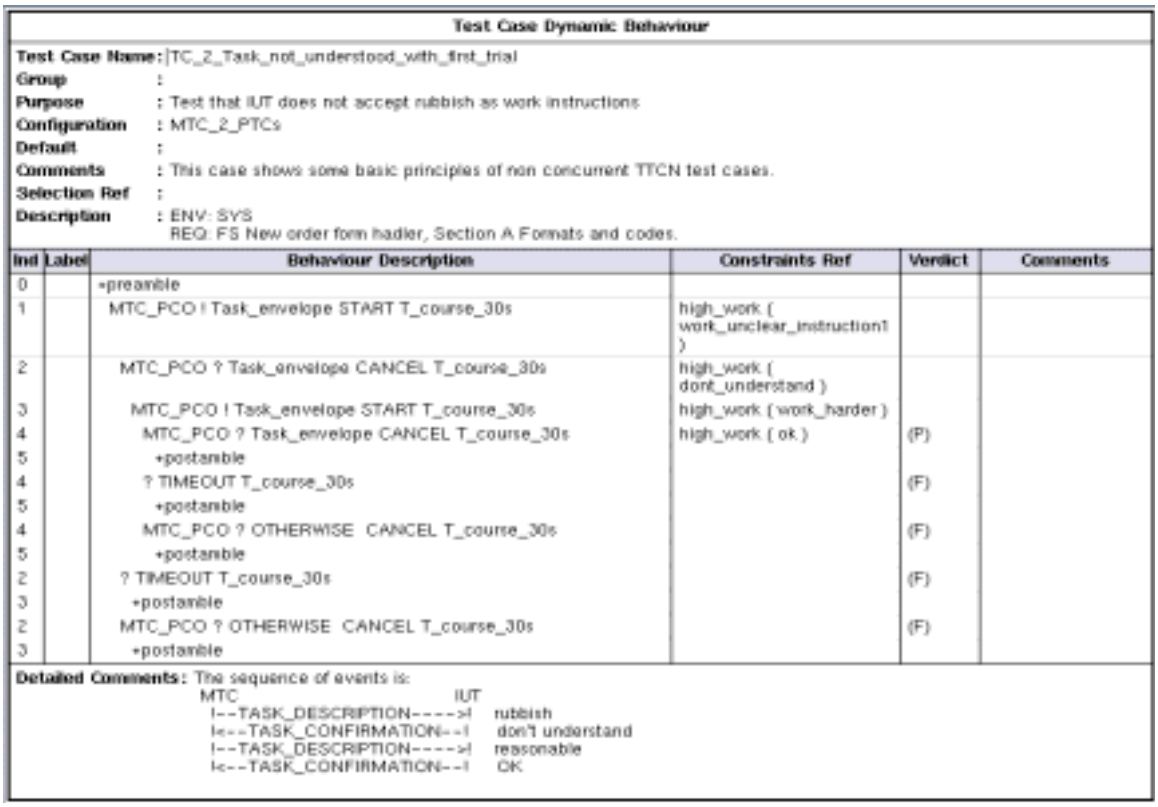
TTCN-2

TTCN (Tree and Tabular Combined Notation) was first published as a standard by the International Organization for Standardization (ISO) in 1992. Since then, it is more and more used by the big telecommunication vendors as well as the operators and test laboratories. The language is supported by a large variety of sophisticated tools such as test systems, editors, compiliers, syntax checkers and simulators. TTCN has been used successfully to specify tests mainly for conformance testing but it is used also for functinal testing. It has been called TTCN-2 since 2000 when the new version of the language TTCN-3 had been launched.

TTCN can be written in either of two notations, one of them is a graphically enhanced form for the human reader, TTCN.GR, and the other is a machine processable form, TTCN.MP. A test suite in TTCN.GR format is a set of tables. An editor supporting the reading and writing of the tables is extremely necessary for the users to navigate through the tables.

A TTCN test suite consists of five parts:

- Overview part describes the hierarchical group structure of the test suite and lists with comments the test cases, test step and default behaviour.
- Imports part which identify object imported for reuse. The test suite is built of modules and objects can be reused in modules other than where it was defined.
- Declarations partin which data types and other objects relevant for the test suite are declared, data objects can also be defined in ASN.1.
- Constraints part in which the concrete data objects and patterns for data object associated with send and receive event are defined,
- Dynamic part in which the test cases that implement the test purposes are described (Figure 3.10.2). This part contains also test steps, that are common parts of test cases collected in a specific library in the dynamic part, and the default library in which parts of the behaviour descriptions common to several test cases are collected.



Test Case Dynamic Behaviour

Test Case Name: TC_2_Task_not_understood_with_first_trial
Group :
Purpose : Test that IUT does not accept rubbish as work instructions
Configuration : MTC_2_PTCs
Default :
Comments : This case shows some basic principles of non concurrent TTCN test cases.
Selection Ref :
Description : ENV: SYS
REQ: FS New order form hadler, Section A Formats and codes.

Ind	Label	Behaviour Description	Constraints Ref	Verdict	Comments
0		+preamble			
1		MTC_PCO ! Task_envelope START T_course_30s	high_work (work_unclear_instruction1)		
2		MTC_PCO ? Task_envelope CANCEL T_course_30s	high_work (dont_understand)		
3		MTC_PCO ! Task_envelope START T_course_30s	high_work (work_harder)		
4		MTC_PCO ? Task_envelope CANCEL T_course_30s	high_work (ok)	(P)	
5		+postamble			
4		? TIMEOUT T_course_30s		(F)	
5		+postamble			
4		MTC_PCO ? OTHERWISE CANCEL T_course_30s		(F)	
5		+postamble			
2		? TIMEOUT T_course_30s		(F)	
3		+postamble			
2		MTC_PCO ? OTHERWISE CANCEL T_course_30s		(F)	
3		+postamble			

Detailed Comments: The sequence of events is:

MTC		IUT
	!--TASK_DESCRIPTION----->!	rubbish
	!<--TASK_CONFIRMATION--!	don't understand
	!--TASK_DESCRIPTION----->!	reasonable
	!<--TASK_CONFIRMATION--!	OK

Figure 3.10.2 Dynamic behaviour table in TTCN-2

The Test case dynamic behaviour table on Figure 3.10.2 has a header part, as all other tables in TTCN-2 have. This contains the name of the test case, the reference to the test purpose the test case implements and finally it is also possible to define a selection reference for the test case. Depending on this reference the test case will be selected for execution during the test campaign. The body of the table contains the behaviour description coloumn where the messages (PDUs, Protocol data Units, and ASPs Abstract Service Primitives) to be sent and expected as answers are given. The lines are indented which defines the sequence of the events. The more indented message comes after the less indented ones. The same indentation means an alternative event therefore it can be only received messages indented the same level. The ! mark stands in front of the messages to be sent by the tester and the ? mark is placed in front of the received messages.

The constraints Reference coloumn contains the references to the tables that defines the data, the parameter values in the messages, these are the tables in the Contsraints part of the test suite. The data types of messages can be given either in a tabular definition format or in ASN.1. ASN.1 is used in the test suite for those protocols which protocol standard contains ASN.1 data definitions. The

ASN.1definitions are the same in the test suite as in the protocol standard just they are put in tables. ASN.1 is more and more used in protocol design, and so in conformance testing. The verdict coloumn gives the final verdict for the alternative incoming events defined in the behaviour part. Only one pass verdict can be given in one test case.

The execution order of the test cases in a test suite is not given. Test cases are logically grouped in functional test groups that help the systematical writing and the reading of the test suite. The main building block of a test suite is the test case, and in a test case the elementary units are called test events. These are usually exchange of messages like PDU (Protocol Data Units) and ASPs (Absract Service Primitives) at given PCOs. Common parts of test cases are put into test steps, that can contain other test steps for any depth of nesting.

TTCN-3

There was a growing need for a test language that is applicable beyond the relatively narrow field of conformance testing. The new version of TTCN, which became known as TTCN-3 was published in 2000.

From a syntactic point of view, TTCN-3 differs from earlier versions of the language, it is a modern programming language. It has many new features, but it also retains much of the proven functionality of TTCN-2. TTCN-3 has the look and feel of a modern programming language. Test specifiers and engineer will find this general-purpose, flexible and user-friendly test language easier to learn, easier to use and easier to implement.

Its new syntax and additional capabilities should attract users from technologies not normally associated with TTCN. Because it is closer to the languages used in product implementation, TTCN-3 will be more acceptable as an integral part of manufacturers' own development process.

TTCN-3 can be used for protocol testing for mobile and Internet protocols, supplementary service testing, module testing, the testing of CORBA-based platforms, the testing of Application Programming Interfaces (APIs) and many more applications. The language is not restricted to conformance testing, but can be used for interoperability, robustness, regression, syste, and integration testing.

The syntax of TTCN-3 may be new, but the language has retained (and improved upon) much of the well proven capabilities of its predecessors. Its main features include:

- Dynamic, concurrent testing configurations
- Synchronous and asynchronous communication mechanisms
- Encoding information and other attributes (including user extensibility)
- Data and signature templates with powerful matching mechanisms
- Type and value parameterization
- Assignment and handling of test verdicts
- Test suite parameterization and test case selection mechanisms
- Combined use of TTCN-3 with ASN.1 (and potential use with other languages such as IDL)

TTCN-3 has a well defined syntax, interchange format and static semantics. Optional presentation formats (eg. tabular conformance presentation format, Graphical presentation format) are defined to make it easier to the user to read a test case written in TTCN-3. The developers of TTCN-3 were well aware of the considerable investment already made in TTCN. So TTCN-3 includes an optional presentation format that caters for conformance testers by displaying test specifications in the familiar tabular format.

The test-based Core Language will be the natural choice for those who are used to conventional programming environment. For users who prefer to express tests in the form of MSCs, this display format is also possible.

3.10.5. Test configurations and test execution

Test configuration defines the architecture of the test system and its connection towards the IUT (Implementation Under Test). Test configuration depends very much on what type of testing is executed. In conformance testing there are only the tester and the IUT that is part of the test configuration. The tester equipment can be connected to the IUT on several interfaces with its paralel test components (Figure 3.10.3). The main test component coordinates the test execution and several test components can send and receive messages in the same time to and from the IUT. These messages pass the PCO (Point of Control and Observation) that is the abstract interface between the tester and the IUT. The tester usually connects to the

IUT via the underlying service provider to the lower layer boundary of the IUT. Sometimes also the upper layer boundary of the IUT is reachable during testing. The test components of the IUT can also send messages to each other, these are called coordination messages and sent on the CPs (Coordination Point). The test configuration can be defined in the test suite with both TTCN-2 and TTCN-3. Test cases are written for given test configurations so it is important to define the configuration together with the test case.

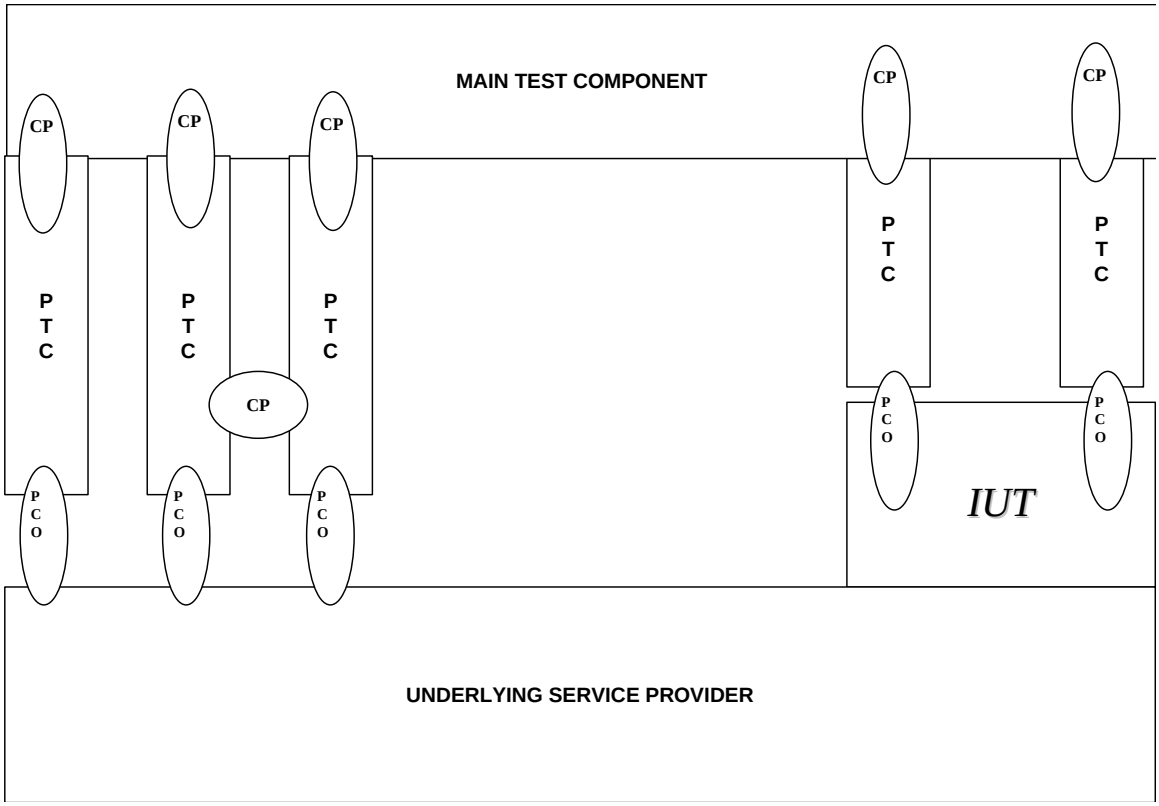


Figure3.10.3 Test configuration

Test execution is the main part of the test campaign, but not necessary the most time consuming. First of all a test equipment is needed that is able to execute TTCN test suites. Also the test suite has to be written, ar chosen if there is standardized test suite available for the protocol we are testing. Those test cases have to be selected from the test suite that are concerning not implemented features of the implementation based on the PICS. During the static conformance review it is checked if all mandatory features are impelemnted and will be tested during the test execution. Also the physical test configuration should be built. All these activities should be done before the test can be executed. As conformance testing is automatic the test execution needs very few manual intervention just some further configuration

can be happen at sertain stages of the test execution. Finally the test report is evaluated and the test results are written in the PCTR.

Conformance testing is the most elaborated test method that is strictly standardized. For other software and protocol test activities no standars are available as they play important role in the developoment process of the vendors but not key issues from the point of view of the interoperability of the heterogenous network elements. It is only interoperability testing that is mostly done nowadays on Internet protocol implementations but the interest towards conformance testing is also growing in this community.

References

[3.10.1] Methods for Testing and Specification (MTS);
The Tree and Tabular Combined Notation version 3;
TTCN-3: Core Language
ETSI DES / MTS-00063-1 Version 1.0.10, November 2000

[3.10.2] Methods for Testing and Specification (MTS);
The Tree and Tabular Combined Notation version 3;
TTCN-3: Tabular Presentation Format
ETSI DES / MTS-00063-2 Version 1.0.3, November 2000

[3.10.3] Methods for Testing and Specification (MTS);
The Tree and Tabular Combined Notation version 3;
TTCN-3: Graphical Presentation Format
ETSI DES / MTS-00063-3 Version 1.0.4, November 2000

[3.10.4] B. Baumgarten and A. Giessler, *OSI Conformance Testing Methodology and TTCN,* Elsevier, Amsterdam, 1994

[3.10.5] B.S. Bosik and M.Ü. Uyar, Finite state machine based formal methods in protocol conformance testing: from theory to implementation, *Computer Networks and ISDN Systems* **22**, 7-33, 1991

[3.10.6] ETSI ETR 266; Methods for Testing and Specification (MTS); Test Purpose style guide, 1996

[3.10.7] ITU-T Recommendation X.290-X.296 – ISO / IEC 9646, Information Technology – Open Systems Interconnection – Conformance Testing Methodology and Framework, 1994

[3.10.8] ITU-T Recommendation Z.500, Framework on Formal Methods in Conformance Testing, 1997

[3.10.9] G.J. Holzmann, *Design and Validation of Computer Protocols,* Prentice-Hall, Englewood Cliffs , New Jersey, 1991.

[3.10.10] K.Tarnay, *Protocol Secification and Testing,* Akadémiai Kiadó, Budapest, 1991

[3.10.11] M.Ü. Uyar, Dual-state Augmentation for Minimizing conformance Test Costs, *Computer Networks and ISDN Systems* **30**, 1277-1294, 1998.

[3.10.12] A. Cavalli, Different Approaches to Protocol and Service Testing, *Testing of Communicating Systems, Methods and Applications* Csopaki,Dibuz, Tarnay (ed.), Kluwer Academic Publ. 1999. Pp. 3-18.

[3.10.13] T.Csondes,S.Dibuz and P.Kremer: Experiments on IPv6 testing: *Testing Of Communicating Systems, Tools and Techniques*, Ural, Probert, Bochmann (ed.)

Kluwer Academic Publ., 2000. pp. 113-126.

[3.10.14] Edward Kit: Software Testing in the real world Addison Wesley, 1995.

4. Networks

Telecommunication networks are established by combining transmission and switching functions, interconnecting various telecommunication devices. A number of connections are simultaneously set up over the telecommunication networks, which are bearing various telecommunication services. Connections can be unidirectional and bidirectional. The type, size and capacity of the telecommunication networks are determined by the quantity and geographic distribution of the users connected to the network and the complexity, frequency bandwidth, traffic features and quality expectation of the demanded services. According to this wide meaning, this chapter discusses the program distribution and broadcasting, too. (Therefore such a set of networks is recently called electronic communication networks.)

The architecture of the telecommunication networks significantly changed due to the pervasive evolution of the network devices and technologies and the huge growth of the demands. An enormous increase of the network intelligence and the bandwidth available to connections has been and will be presented. At the beginning specific networks, as telephone networks, data networks, program distribution networks have been established, fitted to the type of the information to be delivered. By enlargement of the technological opportunities and widening of the variety of services, integrated network solutions are emerged, as ISDN, ATM and IP networks, which are able to multi-service provision and higher network utilisation. In the future evolution of network architectures the mobility, the network access for mobile users plays more and more important role in addition to the bandwidth and network intelligence.

In the course of the telecommunication network planning the considered service demands (variety and quality of the services, their traffic volume and distribution changing in time) are to be satisfied by network devices and technologies available such a way that we take into account:

- the existing network, its configuration, capacity, capability to expansion and development or its obsoleteness, replaceability;
- the installation limits, i.e. geographic (terrain, hydrographic, urbanistic) or other constraints deriving from any agreement;

- the technical, quality, reliability and survivability (security) requirements against the networks, which are described by technical standards, regulations, specifications and agreements;
- the economic aspects, the limits of the sources, the schedule of their availability, the possible stepping of the extension;
- the expected evolution of the network devices and technologies, the changes of the costs, the emergence of the innovative technical solutions, as well as
- uncertainties of the forecast of the service demands, the deviation of the traffic flows, the launching of newer services, etc.

The network solution resulted from the planning process includes the configuration and capacity parameters of the network, as well as the scheduling of its development. In case of economic optimisation the network solution is chosen out of the network solutions corresponding to the above-mentioned conditions taking into account the investment and operational costs emerging in a given time horizon. Typically, the solution meeting the demands with minimum costs is considered the optimum solution. In case of insufficient resources, optimisation is carried out for maximising the satisfaction of the demands. In certain cases the maximum of the demand-to-cost ratio, the solution with maximum efficiency is searched. Choosing the best network solution there are other, investment protection aspects to be taken into account in addition to the economic criteria, as the coexistence of the existing and new network technologies, the flexibility of the new network platform for adopting unforeseen services, the emergence of a subsequent network platform supporting a wider integration of telecommunication services

The Chapter is divided into two parts. The first part of the Chapter consists of three subchapters which overview the common knowledge related to the telecommunication networks, including the classification of the networks, the layer model of the networks, the economic, reliability, survivability and other requirements of the network architecture planning, as well as the principles of the procedures widely-used in the different purpose planning.

The second part of the Chapter is devoted to the various types of networks, it embraces nine subchapters. In the first three subchapters the fixed, voice-based networks (telephone network, ISDN), the different forms of datacommunications (mentioning telex network, too) and the traditional wire-line technological networks are discussed. Then the program transport and broadcasting, and the cable-television networks are presented. The subchapter devoted to the terrestrial cellular

mobile networks involves not only GSM and 3-generation mobile (UMTS) networks, but paging, cordless and trunked radio systems, too. After the discussion of the satellite networks and the traditional IP networks, the closing subchapter presents the real-time IP networks, as an appropriate architecture – according to our present knowledge – is to deploy the infocommunication convergence.

Gyula Sallai, DSc, Editor of the Chapter

4.1. Network Structures

Henk Tamás dr., author

Gyula Sallai dr., reviewer

The *network structures* subsection summarises the general knowledge in networks, including the definition of the networks, their classification and development trends, the basics of routing, and functional and topological network models. During the description we mostly rely on the seven-layer OSI reference model (Open System Interconnection), described in Subsection 1.10, which we will refer in short as OSI model. We describe the network structures in the third layer of the OSI model, which is called the *network layer*.

4.1.1. Network Definitions

The purpose of the networks is the transfer of *different types of information* (e.g., voice, sound, document, textual or multimedia message, video data, message-stream data, byte-stream data, interactive data) either separately or in an integrated way. *Different networks* have been developed corresponding to the transmitted information types providing *different services*, e.g., telephony-service network, or shortly telephony network; integrated services network or shortly integrated network. Networks providing different services, when realised in practice, many times may partly share their building blocks, but they can be theoretically defined separately based on the provided service (e.g., telephony network, leased-line network). The networks develop, change, converge, and altogether they are traditionally named as telecommunication networks, recently called as electronic communication networks. In the followings we will use the term *network* in its widest sense.

The service provided by a network can basically have two types, *bearer service* or *teleservice*. The *bearer service* means electronic information transmission without terminals and applications, e.g., 64 kbit/s transparent data transmission. The bearer service is provided by the *bearer network*. The *teleservice* includes the *technical application* provided by the *terminal*, e.g., telephony or telefax service. The technical application may support different *social applications* (e.g., public telephony

network, distant learning), utilised by the *user*. The user may either be a human being or a machine.

Networks can be connected in a peer-to-peer flat structure and/or hierarchically. Thus we got *interconnected networks*, consisting of *basic networks*, and interworking *units* (or *gateways)* connecting them. The basic networks may differ in their technology, area, management and operation units.

In *peer-to-peer interoperating networks* the basic networks provide either only teleservice or only bearer service. Furthermore, the services provided by each must have common elements, and a part or all of these common elements are provided by the interconnected network. The sketch of the interconnection is simple (Figure. 4.1.1). As an example, see the interconnection of two local telephony networks that are using either the same or different technologies, in order to cover a wider geographical area.

In the case of *hierarchical interoperating networks* the bearer network provides a bearer service to the other, *superimposed network*, which provides bearer or teleservice. The parts of the superimposed network surround the bearer network in the sketch of the interconnection (Figure. 4.1.2.a). Networks can be superimposed on each other several times. The interconnected network built of bearer and superimposed networks is called a *network architecture* or *network stack*, consisting of *network layers*. The network layers may differ from technology, area, management or operation point of view; unlike the OSI model layers, e.g., the network layer, which differ functionally from each other (Subsection 4.1.6). The topmost network layer is connected to terminals, thus it provides teleservice.

The topmost network layer may in one or both sides degenerate to a terminal (Figure. 4.1.2.b). The interworking unit can be called in this case a terminal adapter unit. E.g., using secondary data transmission over a bearer telecommunication network using modems, as terminal adapter units. The service set available to the user is determined only by the topmost network layer, although the quality of the service can be influenced by the lower network layers, too. A layered model can as well be set up for the network stack (Figure. 4.1.2.c). This way of modelling emphasises that during the information transmission an upper network layer uses the services provided by the network layer below it.

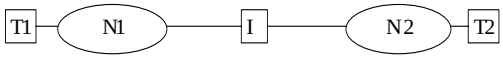


Figure. 4.1.1. Peer-to-peer interoperating networks

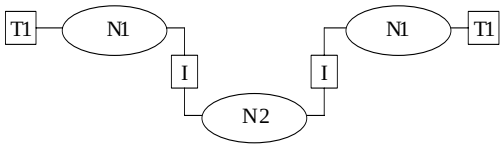


Figure. 4.1.2.a Hierarchical interoperating networks

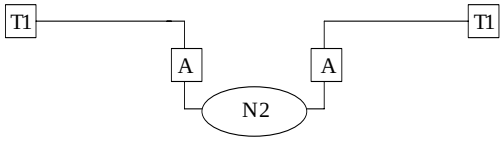


Figure. 4.1.2.b Hierarchical interoperating networks, where network H1 has been degenerated to a terminal

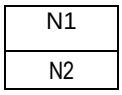


Figure. 4.1.2.c Layered model of a network stack

Legend: Nn: networks I: interworking units

Tn: terminals A: terminal adapter unit

In the network layer the networks can be modelled with nodes and links, which transmit the information between the nodes. This model of the network is sometimes called the *logical* or *logical network*. The links are realised by the lower layers. In the physical layer the signal is propagated in a directed medium (non-branching wire, directed terrestrial or satellite radio) or in a shared medium (bus, broadcasted terrestrial or satellite radio). A medium is shared if the signal of the transmitter physically reaches many receivers. This model of the physical layer is sometimes called the *physical network*.

Taking into account the possible terminals connected to the network and the interconnecting capability of the networks, the networks consist of the following elements:

- *nodes*, with their inputs and outputs,
- uni- or bidirectional links
- network terminals
  - terminal equipments, only in the case of teleservices
  - terminal adapter units, if the technology of the terminal and the rest of the network differs,
  - interworking units or gateways, for interconnecting networks

- *network connection points*, connecting the network terminals to the rest of the network.

The network connection point is connected to the *boundary node* of the network. The rest of the network nodes are called *interior nodes.* The terminals, the boundary nodes connected to the terminals, and the network parts between these two elements are together called as *access network*. The boundary nodes, the internal nodes and the network parts between them are collectively named as *core network*.

The node may do the following operations with a piece of information arriving at one of its input ports:

- store, delete, modify
- forward
  - on one outgoing port (unicast node)
  - on several outgoing ports (multicast node)
  - on all of its outgoing ports (broadcast node)

Unicast and multicast nodes have the *routing* capability. The network wide routing capability of the network nodes are called as *routing*.

The information transmission capability of the network nodes and links are determined by their capacity, e.g., number of switching points, bandwidth, etc. If the quantity of information reaching a node or link exceeds the capacity of the node or the link, then the network becomes *congested*. To avoid a congestion or to recover from a congested state the network has the *congestion protection* capability. The two basic tasks of the network layer of the OSI model is the routing and the congestion protection.

The method of routing and congestion protection depends on whether during the information transmission a circuit is built up or if segmentation into packets occurs. In the case of *real circuit based* networks a real circuit is built up between the terminals and the information is not segmented into packets. The real circuit means fixed path in the network during whole time of the call or lease, which can only be changed due to the movement of the terminals (in mobile networks). Several (2 – 100 000 – 20 000 000) circuits go through a node or link via the *multiplexing of physical channels*. The real circuit is realised in the physical and data link layers, but from

routing and congestion protection point of view it is defined in the network layer, too. In the case of *packet based* networks the information is segmented into packets, and there are two ways of forwarding the packets: either to pin the route of the packets *for the whole time* of the call or lease, thus virtually creating a circuit, a so-called *virtual circuit* in the network layer; or not to fix the route of the packets for the whole time of the call or lease, that is not creating a circuit in the network layer, not even virtually. This latter case is called as *in the network-layer circuit-free network*, or shortly *circuit-free network*. The electric signals in the physical layer are transmitted by circuits, physical signal circuits, in this case, too. Based on these, the networks can be categorised as:

- real circuit based networks (circuit based, not packet based)
- virtual circuit based networks (circuit based, packet based)
- circuit-free networks (not circuit based, packet based)

4.1.2. Network Categorisation

The categorisation of the electronic communication networks is shown in Table 4.1.1, according to the current development level of the networks. These networks are explained in detail in Subsections 4.4.-4.12. The categorisation could be done several ways; the classification in Table 4.1.1 is based on the groupings of the routing and congestion protection capabilities of the networks, as described in Section 4.1.1.

The *broadcast sharing networks* provide multicast bearer service to the broadcast radio network senders and to the cable-TV feeding. In *traditional broadcast communication networks* (analogue broadcasting or cable-TV network) the same information can be received at every access point, i.e., these networks provide broadcasting teleservice. In *interactive broadcast communication networks* (digital programme sharing and broadcast radio network, interactive cable-TV network) supplementary routing capability is added to the main broadcasting feature.

The *infocommunication networks* operate by routing, or mainly by routing and additionally by broadcasting. Following the common terminology, the circuit-based networks in the rest of Subsections 4.1 and 4.2 are referred shortly as *telecommunication networks*. The telecommunication networks guarantee the quality

Electronic Communication Networks																					
Broadcast communication networks						Infocommunication networks															
Broadcast Sharing Networks		Broad-cast Radio Networks		Broadcast Distributi-on Networks		Circuit-based / Telecommunication Networks											Circuit-free / Computer-Networks				
						Narrowband Telecommunication Networks							Broadband Telecom-munication Networks								
analogue broadcast sharing network	digitalis broadcast sharing network	analogue broadcast radio network	digital broadcast radio network	cable-TV network	interactive cable-TV network	Telex network	telephone network	leased-line networks	circuit-switched data network	packet-switched data network (X.25)	(N-)ISDN network	mobile communication networks (1G, 2G, etc.)	Frame Relay (FR) network	ATM network	SDH network	optical network	Internet network	QoS IP networks	mobile Internet networks (3G; WLAN, etc.)	other computer- networks	

Table 4.1.1. Categorisation of Electronic Communication Networks

of the service due to their circuit-based nature – depending on the design of the network structure. Moreover, they realise a well scalable medium sharing, in terms of geographic location and bandwidth.

The *narrowband telecommunication networks* are mainly realised by the real circuit and PCM (Pulse Code Modulation) based PDH (Plesiochronous Digital Hierarchy) and by the virtual circuit based X.25 technology. These technologies are capable of providing both teleservice and bearer service. Among the narrowband telecommunication networks shown in the table, networks providing each of the two types of services are listed assorted. In narrowband telecommunication networks providing teleservice the terminal is usually simple and cheap, which enables the widespread usage. The simple terminal only allows a couple of services within one service class. The intelligence is realised mainly in the network. The access network is typically analogue, the core network is digital. The narrowband telecommunication networks operate up to the following upper speed limits: in fixed access networks 2 Bit/s, in mobile access networks 64 kbit/s, in metal-wired core networks 140 Mbit/s.

Among the *broadband telecommunication networks* the virtual circuit based FR (Frame Relay) and ATM (Asynchronous Transfer Mode) networks could theoretically provide teleservice as well, in the practice, however, they only provide bearer service. The real circuit based SDH (Synchronous Digital Hierarchy) and the optical networks are only capable of providing bearer service. Therefor the naming of the broadband telecommunication networks refer to the used technology.

Computer networks are designed and optimised for information exchange between computers. Computer networks are packed based and are not using virtual circuits. The complex operation mechanisms of the terminal can fundamentally contribute to the establishment of the quality of service of the networks, e.g. via flow control. Nevertheless, no or at most only relative guarantee can be provided to the service parameters. The access speed of the router nodes is well scaleable: this speed covers a range of several orders of magnitudes (1 kbit/s – 10 Gbit/s). Although the elements of the network are often not reliable enough, the dynamic feature of the routing algorithms and the intelligence of the terminal together result in a reliable transmission. The most widely used computer network is the Internet, which is based on the IP (Internet Protocol) protocol family. The most important elements of the family are: Ethernet or PPP (Point-to-Point Protocol) access, MPLS (Multi-Protocol Label Switching) bearer network with label switching, IP based routing, TCP (Transmission Control Protocol) transmission with flow control or UDP (User Datagram Protocol) transmission without flow control.

On one hand, in the telecommunication industry the computer networks are often simply considered as telecommunication data networks. On the other hand, in the computer industry the telecommunication networks are often not mentioned at all, or are simply considered as a part of the computer network, which realises the physical interconnection. The approach of Subsection 4.1 is a balance of the two described approaches. Accordingly, the computer networks are mostly built on top of bearer telecommunication networks using the already built telecommunication networks: e.g., over bearer telecommunication core network in TCP/IP/ATM/SDH structure, or over bearer telephony network with modem adapter unit, etc. Secondly, a telecommunication network can also be built on top of a computer network, e.g., voice transmission over bearer IP network. Thirdly, equal co-operation is also possible between telecommunication and computer networks, e.g., between the WAP

(Wireless Application Protocol) capable GSM (Global System for Mobile communications) terminal and the WAP gateway of the network data network transmission is realised in a circuit based network, and then between the WAP gateway and the WWW (World Wide Web) server a computer network is operating. The computer network can be built on top of an interactive broadcast communication network, and vice versa.

Looking back at the development of the infocommunication networks, we can notice that the routing and congestion protection technology had a spiral way of development: message switching in telegraph networks[1], real circuit switching in telephone networks, virtual circuit switching in data networks, network layer routing in IP networks, dynamic route handling with label switching in MPLS networks (see 4.1.3.). A similar development spiral can be observed in the transmission medium sharing technology, too: space division, frequency division, and time division e.g. in the telephone network, and later in the optical networks, too. The spiral way of development can be interpreted for the principle of the routing algorithms (Section 4.1.4.) and for the network forms of the topology models (Section 4.1.7.), too. When they were introduced certain technologies and principles many times belonged to networks transmitting certain information types, then they extended to secondary transmission of different information types, or to networks serving for different information types, and practically to integrated networks transmitting several information types simultaneously. This process is considered as network convergence. The process of the development and convergence is influenced by economical and profession-political factors.

4.1.3. Elements of Routing

[1] The telegraph network realised a manual message switching. This means, that the information was transmitted encapsulated in message units, so that the message was re-telegraphed from node to node. The storage of the message in a node was done in the form of Morse codes written on a paper tape. During transmission part of the message was still stored on the paper strip at the sender side, the other part was already on the paper tape at the receiver side. In every node it was decided which way to forward the message. The telegraph network was the first electronic telecommunication network, which used the “store and forward” principle – on message basis. The Telex network first operated by message switching, too, but the paper tape with the Morse codes has been replaced by the punched tape. Later on the operation principle of the Teletype networks has been changed to circuit switching, taking the approach of the telephone networks. The “store and forward” principle is nowadays used in the packet based networks, using much faster storage units.

Type of Routing			User Control			Network Management Control		
			Routing Name	Routing Device	E.g.	Routing Name	Routing Device	E.g.
circuit based		real circuit based	call routing	circuit switch	PDH	circuit cross connect	digital cross connect	SDH
	packet based	virtual circuit based			ATM			ATM
		circuit free	packet routing	router	IP	path routing	label switch	MPLS

Table 4.1.2. Elements of the Routing

In infocommunication networks the traffic is controlled by solving the problem of routing in the nodes of the network. The possible ways and means of the solution of routing is summarised in Table 4.1.2. The notations are defined partly by the logic of the table and partly by the international terminology.

The network manager – just as the user – can either be a human being or a machine. The task of the network manager is to group and reserve on demand the capacities of the nodes and links of the network or the basic network in the following networks:

- core network of public or private switched networks
- core and access network of public or private leased-line networks
- virtual private networks

The routing can happen in the following ways:

- the user initiates a call and the routing algorithm of the network builds up the path of the circuit for the time of the call (call routing)
- the user sends a packet with individual addressing and the routing algorithm of the network decides the path of the packet for each packet individually (packet routing)
- the network manager creates circuit trunks or circuits for a longer period, such as several months or years (circuit cross connect)
- the network manager creates a virtual subnetwork where the packets of certain sender-receiver couples travel on designated paths, which paths, however, may dynamically be rearranged even during the sending of the packets (path routing).

A node can act both as a circuit switch and as a digital cross connect simultaneously. E.g., in ATM networks not only virtual circuits are defined between the access points of the network, but also virtual paths are defined for the virtual

circuit trunks, which travels between the nodes of the networks or in a subnetwork within the ATM network.

If the number of the input circuits of a circuit switch is higher than the number of its output circuits then the circuit switch does a *traffic concentration.* In this case the set-up of a circuit may be blocked in the network due to call congestion. Similarly, if the sum of the input bandwidths of a router is higher than the sum of its output bandwidths then the router is a traffic concentrating router. The traffic concentration may result in packet congestion in this case, which results in degraded quality of the service. The utilisation of the network improves in turn. In real circuit based controllable digital cross connect there is no traffic concentration possibility, i.e., the sum of the input bandwidths equals to the sum of the output bandwidths. Nevertheless, in the case of virtual circuit based controllable digital cross connects and label switches there is traffic concentration possibility as well.

If the network contains circuit switches, then it is *circuit switched*, or shortly *switched network.* Switched networks may contain digital cross connects. The nodes of *leased networks* providing guaranteed bandwidth and service are digital cross connects, in these networks there cannot be circuit switches, routers, or traffic concentrating digital cross connects or label switches. Packet based leased networks providing bandwidth controlled by charging and limited quality of service may contain traffic concentrating digital cross connects, traffic concentrating routers and/or traffic concentrating label switches as well, e.g., *virtual leased network, virtual private network*.

In network stacks, e.g., IP/ATM/SDH networks the routing, the virtual circuit/path based cross connecting and the circuit based cross connecting might coexist. This way the good network utilisation, the network reliability and the possibility of allocation of resources to the elastic traffic is realised in the network layers separately. Figure 4.1.3. shows a PDH/SDH telephone network.

The four SDH digital cross connects in the middle of Figure 4.1.3. are controlled by the network manager, thus in these devices the paths of the circuits are fixed for long terms, e.g., for months. The capacity of the SDH digital cross connect is fairly high, e.g., approximately 30,000 PCM circuits belong to a 2.4 Gbit/s SDH link. The four digital cross connects form a *bearer core network*. The label switch would be part of this, too. The bearer network is surrounded by PDH circuit switches.

Together with them we call it a *switched core network*. The routers would be part of it, too. If we would not like to emphasise the difference between the bearer and the switched core network, then we can shortly speak about a *core network*.

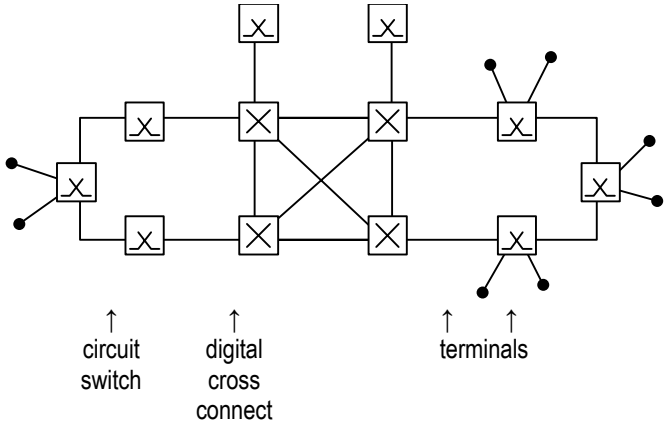


Figure 4.1.3. Structure of a Switched PDH/SDH Network

4.1.4. Principles of Routing

The routing algorithms of the routing capable network layer nodes (such as: circuit switches, digital cross connects, routers or label switches, but a broadcast node is not like this), shortly nodes later on in this section, are determined by the following parameters:

- the number of the nodes in the network
- the topology of the network
- the reliability of the nodes and links of the network
- the maximal time interval, to which the traffic statistics can be well estimated
- special service requirements (e.g., mobility, multicast, etc.)
- the feasibility of the algorithms with respect to complexity and time consumption requirements at the given level of the technological development

In networks consisting of few nodes *democratic routing* is possible: all of the nodes can be equal at the route selection. During the selection of the routes there is no such constraint that it has to go through a certain node. The selection of the routes only depends on the traffic load of the links and nodes. Unrelated to this, the routing algorithm can be centralised or decentralised. In the case of *centralised routing algorithm* the routing centre distributes the routes and alternative routes based on preliminary or dynamic traffic estimation. In the case of *decentralised*

routing algorithm the router nodes are not concentrating the routing capabilities, but notify each other about the availability of their neighbours and about the traffic load of the nodes and links.

With high number of nodes – over 5-200 nodes, depending on the technology –, however, the size of the routing tables radically grow. Therefore both in telecommunication and in computer networks some kind of *hierarchical routing* is usually used. We can get to the basic model of the hierarchical routing if we divide the nodes of the network into groups and assign a leading node in each group. Every node is connected to and only to the leading node with a direct link. In the case of second-order hierarchy, we group the leading nodes as well, etc. The basic model of the higher hierarchical routing eventually follows a *hierarchical tree structure*, where the elements of the structure are the network nodes. The nodes of the same level form the *hierarchical network planes*. The route through the hierarchy is unambiguous, thus it is a called *direct routing*. This network, however, is vulnerable, as the fault of a high ranked link or node results in the segmentation of the network. Reliable hierarchical routing can be realised e.g. by a *hierarchical double tree structure*: two nodes are used instead of one in the root, and every node is connected to two higher level nodes. If the basic model is not fully built, then the tree structure is truncated from the direction of the root and the nodes of the highest hierarchy plane are connected by a topology which is at least double connected. Another theoretical possibility for the realisation of a reliable hierarchical routing is the *structure of the hierarchical rings*: the nodes of a certain group are connected by a ring topology, then two leader nodes are selected from every first-level ring and these are connected by a second-level ring, etc.

In the practice the flat and hierarchic, and the centralised and decentralised routing models are used simultaneously. During the development of the different infocommunication networks the convergence is noticeable in this respect, too. Subsections 4.4. – 4.12. show specific network solutions.

The routing is efficient if, corresponding to the network planes, the *numbering* or *addressing* is hierarchical as well. In each hierarchical plane traffic concentration can happen as well. In the case of complex network stacks, the certain network layers often consist of separate digital cross connects or circuit switches or routers or

label switches. The hierarchy system follows the geographical location and the addition of the traffic in the given area.

4.1.5. Control and Auxiliary Networks

The network nodes are not only parts of the information transmitting network, but *logically separated control networks* are connected to the nodes as well, such as:

- *signalling network*: in circuit based networks it controls the set-up and tear-down of the circuit on the request of the terminal, furthermore in public networks it transmits authentication and charging data,
- *routing network*: operates the routing algorithms of the circuit switches and the routers
- *management network*:- controls the digital cross connects and label switches changes the routing algorithms of the circuit switches and the routers, if needed- statically fills up the routing tables of the circuit switches and the routers, in case of abnormal operation of the routing algorithms of the network. Theoretically all the three control networks may operate the nodes by manual or automatic control. The automatisation of the signalling network was started in 1889. The routing network operates automatically since about 1940, too. The management network is controlled partly automatically, partly manually.

In addition to the network and the control network, several *auxiliary networks* support the operation of the network, e.g., remote supply network, traffic measurement data collecting network, network operating protection switches, etc.

Traditionally the aggregated traffic of the control and auxiliary networks does not exceed the 10% of the network traffic. Nowadays the handling of the various services of the integrated and mobile networks and the supplementary services (e.g., call waiting) heavily increases the traffic of these networks, up to the 50% of the network traffic.

4.1.6. Functional Model of Networks

During the functional modelling of the networks we group the tasks performed by them into hierarchical functional layers. The functional model of the electronic communication networks is not wider than the model of the infocommunication networks, as fewer types of network tasks are carried out in the broadcast communication networks than in the infocommunication networks. We can obtain the

model of the infocommunication networks if we merge the different functional models of the computer networks and the telecommunication networks.

The functional modelling of the network means that we group hierarchically the protocols performing certain network tasks. Such a group is called a (functional) *layer*, with which we build a hierarchic *protocol stack*. We model the information transmission between two network devices, such that there is virtual information transmission between the identical layers of the two network devices. In reality the layers transfer the information hierarchically among each other within one device. In packet based networks, for example, this works the following way: at the sender side going down the hierarchy each layer's protocol data unit is extended with a new packet-header by the layer below it. The actual information transmission between the two network devices happens in the physical medium. At the receiver side going up the hierarchy, every layer removes a header. The certain layers model the operation of the – partly different – parts of the network. The hierarchical modelling emphasises that during the information transmission a higher functional layer utilises the service of the functional layer below it.

The OSI model has originally been developed for modelling computer networks [4.1.8] and [4.1.3]. However, despite the expectations in 1990, the protocols of the seven-layered OSI model have not come into general use in the practice. The realised computer networks are not following the OSI models, and are diverse compared to each other, too. Nevertheless, the OSI model is still the basis of the comparison of the different protocol stacks. In the OSI model the network parts can be assigned to the layers as follows:

- Layer 1-2: between two adjacent nodes of the network, or between the network terminal and the boundary node, or between theterminal and the adapter unit,
- Layer 3: between the network boundary nodes,
- Layer 4-7: between the network terminals.

The protocol data unit in the first three layers has got distinct name, too: bit in the first layer, frame in the second layer, packet in the third. (See Subsection 1.10 for details.)

In Table 4.1.3 we have shown the OSI model, the main layer functions and the five-layered hybrid Internet model [4.1.3], [4.1.4] and [4.1.8]. The telecommunication

networks are described with several layered-models, too. The OSI model has only been adapted to a special case of the telecommunication networks, the data networks within the recommendations of ITU-T (International Telecommunication Union – Telecommunication Standardisation Sector) [4.1.4]. The telecommunication network is described with a four-layered model, the ATM network with a three-layered one, the SDH network with a different three-layered model, etc. In Table 4.1.3 we have presented the five-layered model of the infocommunication networks, proposed by us, which – with practical trade-offs – allows the common modelling of different telecommunication and computer networks. The joint modelling is done such that we compare the main tasks of the different layers and map them to each other, no matter if a network is packet based or not. The cause of the trade-offs during the modelling is that the boundary lines of the functional layered models of the different networks are not exactly aligned.

In the world of computer networks the leased or switched telecommunication network is only considered as a “subnetwork”, which is part of the physical layer. The FR and MPLS protocols, on the other hand, are put in the 2nd layer. In spite of this, we have constructed the layered model of the infocommunication network such that each routing device (circuit switch, router, digital cross connect, label switch) is located to a place corresponding to the network layer of the OSI model. Therefore the transmission layer, which includes the digital cross connect and the label switch, goes up to the level of the network layer. The real circuit switch, the virtual circuit switch and the router is located in the switching layer. We consider the FR switch and the ATM cell switch as type of the virtual circuit switches.

In the layered model of the infocommunication networks the adaptation layer models on one hand the *flow control* between the terminals of the computer networks (cf. TCP in Section 4.11), and models on the other hand the address multiplexing/demultiplexing in the ATM-IP interworking units. The last column of Table 4.1.3 refers to the fact that in the layered model of the infocommunication networks the transfer layer describes the bearer core network and the switching layer describes the switched core network. It is noticeable that the bearer core network consisting of digital cross connects can be very fast and highly reliable because it uses few layers.

In the protocol stack of hierarchical interoperating interconnected networks the network layers (Figure 4.1.2.c) and possibly the data link and the transport layers as well appear multiple times superimposed on each other.

The layered models of Table 4.1.3 refer to the information transmitting networks. For the control and auxiliary networks different functional models can be set up, but in a similar way.

<table>
<tr><th>OSI model</th><th>Layer Functions</th><th>Internet Hybrid Model</th><th>Model of Infocommunication networks</th><th colspan="3">Layer Extension of Information Dispensing Subnetworks</th></tr>
<tr><td>7. Application Layer</td><td>• teleservice</td><td rowspan="2">5. Application Layer</td><td rowspan="2">5. Application Layer</td><td rowspan="5"></td><td rowspan="4"></td><td rowspan="6">access net-work</td></tr>
<tr><td>6. Presenta-tion Layer</td><td>• source coding
• encrypting</td></tr>
<tr><td>5. Session Layer</td><td rowspan="2">• direction handling
• synchronisation</td><td rowspan="2">4. Transport Layer</td><td rowspan="2">4. Adaptation Layer</td></tr>
<tr><td>4. Transport Layer</td><td>• flow control
• (de)multiplexing</td></tr>
<tr><td rowspan="2">3. Network Layer</td><td rowspan="2">• routing
• congestion protection</td><td rowspan="2">3. Network Layer</td><td>3. Switching Layer</td><td rowspan="3">switch-ed core net-work</td></tr>
<tr><td rowspan="2">2. Trans-mission Layer</td><td rowspan="3">bearer core net-work</td></tr>
<tr><td>2. Data Link Layer</td><td>• medium access
• flow control
• error handling</td><td rowspan="2">2. Network Access Layer</td></tr>
<tr><td>1. Physical Layer</td><td>• 2/4 wire conversion
• sender/receiver functions
• signal transmission</td><td>1. Physical Layer</td><td>1. Physical Layer</td></tr>
</table>

4.1.7. Topology Models

The information transmitting networks can be modelled by graphs consisting of nodes and links, too. In the network layer the graph nodes correspond to the network nodes and the graph links to the network links, and the graph is called the traffic topology. The most common topology models are summarised in Table 4.1.4. For the networks in practice a combination of these are used. Among the perspectives used in Table 4.1.4 only the “Bandwidth Load of Links” requires explanation. We consider as connectable nodes the following nodes: the leaves of the tree and double tree topologies, every node except the star nodes in topologies containing one or more star nodes, and further every node in the rest of the topologies. Let us suppose that we have put a traffic source at every connectable node generating a unit of

bandwidth traffic. The question is, how much bandwidth load is put on the links of the certain networks? To answer the question, we examine two extreme cases, the one-to-one connection and the one-to-all connection. In the case of *one-to-one connection* we suppose that every connectable node sends messages to only one other connectable node, such that bi-directional traffic couples are formed. This is the way, for example, how a telephone terminal works in the normal case. In star and tree networks it can be allowed to apply a 2/4 wire conversion in the subscriber line. Therefore in these networks the bandwidth load of the subscriber links are halved. In the case of *one-to-all connection* we suppose that every connectable node sends messages to all of the other connectable nodes. This way the core networks can be simply modelled. During the bandwidth load computation of the links of the star + bi-directional ring topology we supposed that the ring network part carries the traffic of the nodes that are close to each other.

The features listed in the table at certain network forms are typical of the application examples shown in the table. Among the application examples there are two abbreviations that has not been resolved yet [4.1.8]:

- DQDB: Distributed Queue Dual Bus,
- FDDI: Fiber Distributed Data Interface.

Topology	Bus	Tree	Star	Unidirecti-onal Ring	Bi-directi-onal Ring	Double Bus	Double Tree	Double Star	Double Ring	Star + Bi-directional Ring	Full Mesh
Number of Graph Links	n-1	n-1	n	n	n	2n-2	2n-1	2n+1	2n	2n	n(n-1)/2
Bandwidth Load of Links – One-to-one connections	2n	max. n	1	n/2	-	n	max. n/2	-	-	-	-
Bandwidth Load of Links – One-to-all connections	n(n-1)	max. $n^2/2$	2n-2	n(n-1)/2	$n^2/4$ or $(n^2-1)/4$	n(n-1)/2	max. $n^2/4$	n-1	$n^2/8$ or $(n^2-1)/8$	~ 2n+8-$-4\sqrt{(2n+4,25)} \approx n$	2
Network Fault Tolerance	poor					good					very good
Intelligence Requirement in the Connectable Nodes	yes		no	yes							
Requirement for a Node with Distinguished Intelligence	no	in the root	in the star node	no	no	no	in the root	in the star nodes	no	in the star node	no
Supports Data Protection	no		yes	no				yes	no		very much
Propagation Delay	medium			high	medium						minimal
Application Example	Ethernet	hierarchi-cal routing	telephone access network	token ring	SDH network	DQDB network	signalling network	metropoli-tan telephone network	FDDI, SDH networks	metropolitan telephone network	PDH core network

Table 4.1.4. Network Form Models with **n** *Nodes in Infocommunication networks*

References

[4.1.1] Cinkler T., Henk T., Ziegler G. ed.: Távközlő hálózatok (Telecommunication Networks), electronic lecture notes, BME, 2002, http://leda.ttt.bme.hu/~cinkler/TavkHal/

[4.1.2] Balogh T. *et al*: A keskenysávú ISDN kézikönyve (The Handbook of Narrowband ISDN), *Távközlési Könyvkiadó,*Budapest, 1997.

[4.1.3] Hosszú G.: Internetes médiakommunikáció (Internet-based Mediacommunication), *LSI Oktatóközpont,* Budapest, 2001.

[4.1.4] ITU-T: Data Communication Networks: Open Systems Interconnection (OSI) Model

and Notation, Service Definition. Recommendations X.200-X.219.

[4.1.5] S. Keshav: An Engineering Approach to Computer Networking: ATM Networks, the

Internet and the Telephone Network, *Addison-Wesley,* Reading, Massachusetts, US,

1997.

[4.1.6] H.G. Perros: An Introduction to ATM Networks, *John Wiley and Sons*, Chichester, 2001.

[4.1.7] W. Stallings: ISDN and Broadband ISDN, *Maxwell Macmillian International,* New York, 1992.

[4.1.8] A.S. Tanenbaum: Számítógép-hálózatok (Computer Networks), *Panem-Prentice Hall,* Budapest, 1999.

Translated by: Krisztián Németh

4.2. Network Requirements and Principles of Satisfying Them

Tamás Henk dr. (4.2.1-4) and Tibor Cinkler dr. (4.2.5-7), authors

Gyula Sallai dr., reviewer

The *Network Requirements and Principles of Satisfying Them* subsection summarises the requirements set up for the infocommunication networks, including the economical, reliability and usability respects, the traffic and quality demands corresponding to the services. Furthermore it overviews the principles of quality assurance and the technical foundations of the network viability. The terminology of Subsection 4.2 heavily relies on the terms introduced in Subsection 4.1.

4.2.1. Network Structure Requirements

The infocommunication networks are built and operated according to the requirements regarding the Quality of Service (QoS), the reliability, the usability, the viability, the area coverage and the economical operation. In the following we examine each of these requirements.

The main elements of the *quality of service* can be defined as the probability of the following events: the required information transmission can be started, the transmitted information reaches the destination, it arrives with a tolerated error or error-free and further it arrives with a tolerated loss and jitter. The QoS problem is described in detail in Subsections 4.2.2-4. The *reliability* is the reciprocal value of the fault rate of the network. The concepts of *usability* and *availability* take into account the mean time of the fault recovery as well. The fault rate can be kept below an acceptable bound by applying reliable building blocks and with solutions that enhance the *viability* of the network, such as the protection switching and the application of bypass network paths (see Subsections 4.2.5-7). The mean time of the fault recovery can be kept at an acceptable level by storing spare units, organising a standby operation service, and further by contracts with the equipment supplier companies guaranteeing alarm, field-work and fault recovery services with deadlines.

The main elements of the *area coverage* are: the types of the services provided by the network, the geographical distribution of their access points, the fairness between the users regarding the parameters of the quality of service, the reliability and the usability.

The *economical operation* of the network is a complex problem. We can divide the infocommunication networks with respect to their social utilisation into private and public networks. The *private networks* are part of the infrastructure of a company or institution, only authorised users can utilise them up to the level of their authority, but usually without time restrictions and free of charge. The *public networks* can be utilised by any user, who or which can access the network and is able to use the terminal and pays the service charge. Public networks are operated on commercial basis with governmental support or purely on commercial basis. We can speak of the economical operation of the network in the case of public networks operating on commercial basis. The major elements of the economical operation are: the costs of the amortisation and operation of the network, the network utilisation, and the net *profit* based on the charge rates. Complex relations can be observed among the major elements of the economical operation.

The *amortisation* of the network depends on how large network capacity do we create, furthermore on the extent of the specific investment cost of the applied technologies. The network *operation costs* are determined by the automatisation level of the network management, the fault rate of the network elements, the time requirement and cost of the fault recovery and the amount of consumed energy used for the operation. The *network utilisation* is defined by the average ratio of the transmitted traffic and the traffic capacity of the core network links in an hour of normal traffic load. The network utilisation depends on the traffic and QoS demands of the users, the charging policy, the management and reconfiguration possibilities of the network and the network technology. The income based on *charging* may in general depend on the traffic transmitted through the network, the bandwidth and QoS parameters of the traffic and also the distance, the duration and the time of the day of the information transmission.

In the following we are going to overview some of the technical elements of the above mentioned relationships. During this we will characterise the service requirements, summarise the major elements required to fulfil these requirements:

the means of congestion protection and the possibilities of providing QoS, furthermore we will show the relationship between the QoS provisioning and the network utilisation. Afterwards we are going to analyse the possibilities of satisfying the requirements of the network availability and viability.

4.2.2. Categorisation of Service Requirements

We can characterise the certain services available in a given network with:

- the features of the traffic sources, which can utilise the services,
- the quality requirements of the service.

We can classify the services of a certain network according to the above two perspectives into classes, the so-called *service classes*.

Traffic means the information transmission request through a network part or a network originated from one source or aggregated from several sources or from another network. Satisfying a given traffic request is possible following several network conceptions. If the traffic request is present for a higher duration (e.g., several months or weeks) with high duty factor (e.g., in 50% of the time), and the requested QoS is high, too, then it is probably worthwhile to satisfy the traffic request with a leased-line network. In this case the network manager provides a dedicated circuit to the traffic request. If, however, the traffic is present only in a smaller portion of the time and the QoS requirement allows the possibility of call blocking and breakdown, too, then the traffic request can be satisfied with a real circuit switched network. If the given traffic does not require strict real-time transmission, then a packet-based network is satisfactory as well. Encoded traffic sources, like video codecs emit the packets in bursts. To sum up, we can define the traffic coming from a certain source different ways, according to the time scale: the traffic can be a request of a lease contract between two points of the network, a call, a packet burst or a packet.

If a new traffic demand appears in the network then we consider the rest of the traffic as *background traffic* and the following questions has to be answered regarding the QoS:

a.) after the appearance of the new traffic can the QoS requirements of the background traffic be still satisfied?

b.) can the QoS requirements of the new traffic be satisfied with the presence of the background traffic?

To answer these questions the traffic sources and the network mechanisms affecting the QoS parameters have to be modelled. This modelling is fairly complex, especially in *integrated services networks*, where the network is able to satisfy different traffic demands simultaneously. The problem can be simplified if we categorise the service requirements as follows.

The different teleservices and bearer services have different information transmission requirements of the networks. Thus the *traffic parameters* (e.g., call rate, average holding time, average bitrate, maximal bitrate, burst size, minimal sustainable bitrate, speed of the mobile terminal) and the *quality of service requirements* (e.g., delay, jitter, error rate, packet loss probability, packet misdelivery probability, etc.) of the certain services together form the *service requirements*. The service requirements can be categorised and thus grouped into classes. These are the so-called *service classes* of the network. The designing of the network structure and operation is simplified by the categorisation into classes, especially in integrated networks: within a certain service class the same network mechanisms assure the QoS parameters. Such operation principles are described in Subsections 4.2.3-4.

Service classes can be defined in several networks, which were emerged during the historical development of the infocommunication networks. As an example we show in Table 4.2.1 the classes of the packet-based, broadband, integrated, B-ISDN (Broadband – ISDN) infocommunication networks, as defined by the ITU-T. The concept of connection, which appears in the table, will be defined in Subsection 4.2.3.

The service categorisation into classes cannot be considered as a closed problem, just like the network development itself.

ITU-T classes	A	B	C	D
Delay Sensitivity	sensitive		not sensitive	
Bitrate	constant	variable		
Connection	connection-oriented service		connectionless service	

Table 4.2.1. Service Classes in B-ISDN Networks

4.2.3. Congestion Protection

The two major tasks of the network layer of the seven-layered OSI reference model (later: OSI model, see Subsection 1.10) are the routing and the congestion protection (Subsection 4.1). The congestion protection serves as the means of QoS assurance in the network layer. To the QoS assurance, however, the appropriate routing and the other layers (Subsection 4.2.4) contribute as well.

There can be two causes of a *congestion* in the network:

- a certain node receives the information on its inputs with higher speed than it can overall process;
- a certain node would send the information to one of its outputs faster than the maximal speed of the transmission link connected to that output.

In congestion-free networks the *trough put curve* of the network is ideal (Figure 4.2.1). This means, that the aggregated transferred traffic equals to the offered traffic until the transferred traffic reaches the capacity of the network, and after this point the transferred traffic equals to the capacity of the network. In real networks there always exists some level of congestion. With appropriate congestion protection the transfer curve is monotonic increasing, and its limit tangent is parallel with the ideal curve. Without congestion protection the transfer curve might turn back towards the zero transfer. In this case the network collapses. One of the

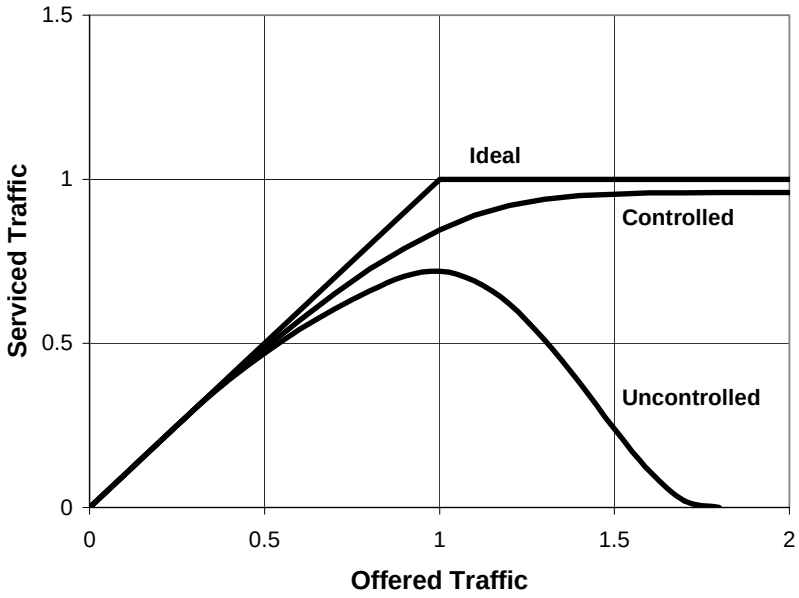


Figure 4.2.1. Typical Transfer Curves

requirements of the network utilisation is that the transfer curve should be monotonic increasing.

The three major elements of the congestion protection are the appropriate routing, the traffic handling and the resource handling. Different network technologies utilise these possibilities to a different extent using different realisation technology.

Routing is built up in practice as a combination of flat and hierarchical routing principles (Subsection 4.1.4). The presence of the flat principle has the consequence that in a certain hierarchical routing network plane in case of congestion on the direct path an alternative path can be chosen. The static alternative routing means that the routing algorithm examines the links of some predefined alternative paths consisting of few links, in a predefined order of the alternative paths. In the case of dynamic alternative routing several (e.g., a hundred) alternative paths can be examined, which can have several links and also the order of the evaluation may depend on the traffic circumstances within the network. With small traffic load the dynamic routing is beneficial with regard to the throughput. In case of significant congestion, however, the network concentrates more and more on searching for alternative paths, and not on the information transmission, if the routing is dynamic. Consequently in case of a significant congestion it is advantageous to switch to static routing.

Traffic handling can be realised with the following means:

- admission control and policing,
- traffic flow handling: flow control, traffic shaping, partial or full traffic dropping.

If we apply *admission control* in the network then the admission is possible only if the conditions a) and b) described in Subsection 4.2.2 are met. The network decides about an admission at the time of the *call* appearance, in real-time. In the case of a positive decision the network and the calling user may contract about the parameters of the traffic source and the QoS and about the charging. In the case of the blo cking of the call, call waiting or call rejection is possible. In the latter case the user may retry to initiate a call. Thus eventually both cases of the call blocking result in a call delay. *Policing* may supplement the admission control: in this case the network verifies if the traffic source exceeds the agreed speed parameters during the operation. If it does, the network may rise the charges and may re-classify the given traffic into a lower service priority, which might even lead to the dropping of the given traffic. *Flow control* means that the congested node signals the upstream nodes to

slow down their information transmission. This mechanism propagates backwards on the path of the information transmission in the network as a chain-reaction, and finally the network might ask the information source, too, to decrease its transmission rate. *Traffic shaping* is the maximisation of the information rate on a given link such that the network stores the congested packets in a buffer allocated for this purpose. *Traffic dropping* might mean the complete dropping of a certain traffic or the partial dropping of its packets.

Resource handling can be realised with the following means:

- resource overprovisioning,
- resource reservation.

In the case of *resource overprovisioning* we build such a level of network capacities that provides that congestion occurs only with a tolerable probability limit according to the estimated traffic parameters. Although overprovisioning is expensive, it is simple and the other congestion protection solutions are not cheap, either.

Resource reservation consists of the following steps:

- selection of path or paths with the sharing of the given traffic load (routing),
- reservation of resources of the nodes and links for the given traffic along the selected path/paths,
- resource deallocation after the cease of the given traffic.

Resource reservation can be initiated by the user in the form of a call. In this case the resource reservation is preceded by an admission control and during normal operation circumstances the resource deallocation is initiated by the calling user, too, with a tear-down initiation. We note, however, that admission control can be applied without resource reservation as well. The resource reservation, on the other hand, can be initiated by the network manager, too, even for an aggregated traffic. Excessive resource reservation requests can be limited using an appropriate charging policy.

Resource reservation means the building of a *connection* in the network layer. If a certain network technology does not build up a connection in the network layer, then the network layer is called *connection-free*. Therefore building a connection is a means of QoS assurance, which guarantees the QoS parameters if the connection is successfully set up. The connection is an abstraction, its realisation in the network

layer is possible by building a fixed path, a circuit (Subsections 4.1.1 and 4.1.3) or a dynamic path (Table 4.2.2). We define the dynamic path such that the path is not fixed for the whole duration of the information transmission, but the path and accordingly the reservation of the resources can be changed during the information transmission, if the traffic load of the network requires it.

Name	real circuit	virtual circuit	dynamic path	data packet based
Connection	connection-oriented			connection-free
Circuit	circuit based		circuit free	
Packet	not packet based	packet based		
Technology, e.g.	PDH, SDH, optical network	X.25, FR, ATM	MPLS, QoS IP	IP

Table 4.2.2. Realisation of Connections in the Network Layer

4.2.4. Comparison of Network Technologies

The service quality is influenced by all the layers of the functional layered model (Subsection 4.1.6). Therefore in each layer it is possible to allocate resources and building a connection is one of the means of doing so. Theoretically each layer of the OSI model can be connection-oriented or connection-free. Building a connection, however, means only in the network layer the setup of a circuit or a dynamic path. In the other layers the connection is realised with different means. Different infocommunication network technologies put an emphasis on the QoS aspects in different layers. The + sign in Table 4.2.3 means that the given layer significantly contributes to the QoS assurance in the particular technology.

In the physical layer the metal-wire technologies have poor error rates (e.g., 10^{-5}), while the optical technologies have good rates (e.g., 10^{-12}). In the data link layer it is possible to achieve a reliable transmission with protection switching or with

Technology	Telephone Network	X.25	ATM	TCP/IP	MPLS, QoS IP
Transport Layer	-	-	-	+	-
Network Layer	+	+	+	-	+
Data Link Layer	+	+	-	-	-
Physical Layer	-	-	+	+	+

Table 4.2.3. Contribution of the Layers to the QoS Assurance

acknowledgements and retransmissions, i.e. with connections realised in the data link layer. In the network layer the congestion protection is the solution for the QoS problem. In the transport layer the *traffic control* may contribute to the QoS provisioning. Traffic control realises a connection between the information source and the information sink in the transport layer. It uses acknowledgements, retransmissions and speed reduction.

Error-free transmission or transmission with low error probability can be realised in every layer. Realising low delay and low jitter transmission, however, is only possible in the three lowest layers. The QoS provisioning methods and manageability features of the different network technologies can be advantageously mixed in the network stack.

Figure 4.2.2 shows the tendency of some of the factors that influence the network profit for some of the network technologies. The network profit is positively influenced by the network utilisation and the available QoS, too, via the charging. The network profit decreases, however, if the algorithms used in the network operation (e.g., congestion protection) are more complex, because in this case the amortisation and the operation cost rise.

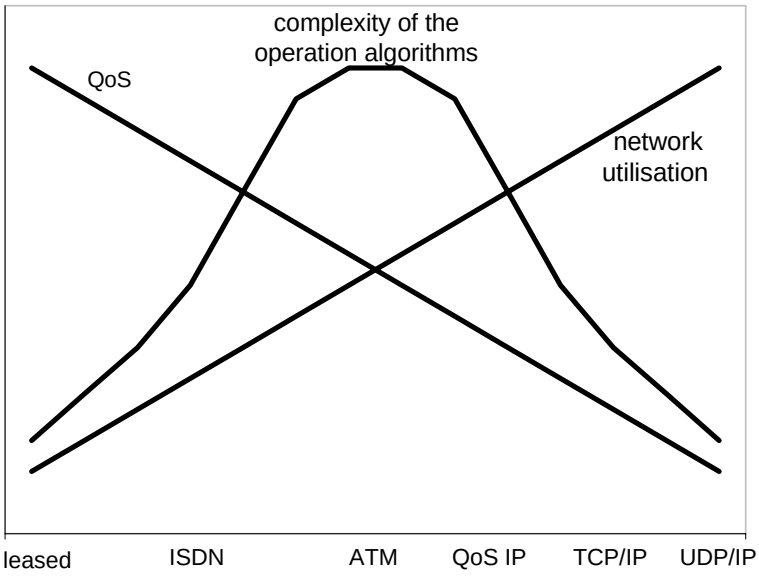


Figure 4.2.2. Factors Influencing the Network Profit

In the figure the network technologies represent (from left to right): the real circuit based leased networks, the real circuit based switched networks, the virtual circuit based switched networks, the dynamic path based networks, the traffic control

based networks and the data packet based networks. Improving the network utilisation generally means the degradation of the QoS. From this respect the trade-off represented by ATM is beneficial. The price of this trade-off is the complexity of the network operation algorithms. From this respect ATM is disadvantageous. That is different network technologies can be advantageous depending on the different user requirements and on their application in a compound network stack.

4.2.5. Network Availability and Viability

The amount of traffic to be carried by modern infocommunication networks grows exponentially. Accordingly, the traffic over certain links of these networks grows similarly. A failure of such a link or other network element leads to traffic loss. On the one hand, this deteriorates the communication of the users, while on the other hand leads to significant losses in revenue for the operator. Therefore, a leased line contract contains not only the bandwidth, duration and price, but also the availability of that line. A customer may ask for availability of "five nines" (99.999%), meaning that during one year the line may be not available for up to 5 minutes. Repairing failed links typically takes hours, or even days. First, the failure has to be detected, then located followed by accessing the exact point of failure before the reparation is really started and completed. Compared to the above 5 minutes this is not allowed! Therefore, the network has to be equipped with resilience mechanisms. The following sections discuss these mechanisms as well as the most important terms and definitions.

Protection or Restoration?

There are in general two ways for enhancing network resilience: the pre-planned static approach and the on-line, i.e., dynamic one. The most significant difference is that while *protection* uses preplanned separated resources for protecting working facilities, *restoration* utilises the resources found free immediately after a failure is detected and localised. To clarify, in case of *restoration* resources are not allocated nor are the protection paths routed, but the decision is taken according the current state of free network resources. Therefore, it can be seen that *protection* requires more network resources than *restoration*, consequently it well be more expensive. However, its operation is faster, and looses less traffic in case of a failure.

Restoration is particularly advantegeous in case of rapidly changing traffic conditions, since it is based on the instant traffic conditions.

There are numerous resilience methods based on both, protection and restoration. The most promising methods are, however, those that utilise both principles jointly. For example two classes of traffic can be distinguished: one that has to be protected, while the other should be restored within the avialable resources. Another method can be to employ restoration with preplanned strategies for speeding up its operation.

Link or Path Protection?

In case of *link protection* the failed link is overbridged only. In Figure 4.2.3 **a** link A-D used by the path A-C fails. In case of link protection only link A-D is bypassed (Figure 4.2.3 **b**), therefore, link C-D carries the traffic twice. Although this method is probably the fastest and simplest one, it has significant resource (particularly capacity) requirements.

Path protection protects any part of a whole path, i.e., if any part of it fails the whole traffic of that path is redirected to the protection path. In Figure 4.2.3.**c** if link A-D fails we switch to path A-B-C. Path protection has two types: path protection *with disjoint paths* and path protection *with failure localisation*.

In case of path protection *with disjoint paths* for each path there is a preplanned protection path which does not contain any common element with the working one except the originating and destination nodes (the two terminating nodes).

In case of path protection *with failure localisation* for each link of a working path there is a protection path, therefore, as many protection paths will be assigned to a working path as many links it contains.Until now we have implicitly assumed that there may be only one failure at time. If we assume that two or more failures can affect the network at the same time, then more than one protection path has to be assigned to each working path. In spite, if restoration is employed it is sufficient to re-run the restoration algorithm after each failure.

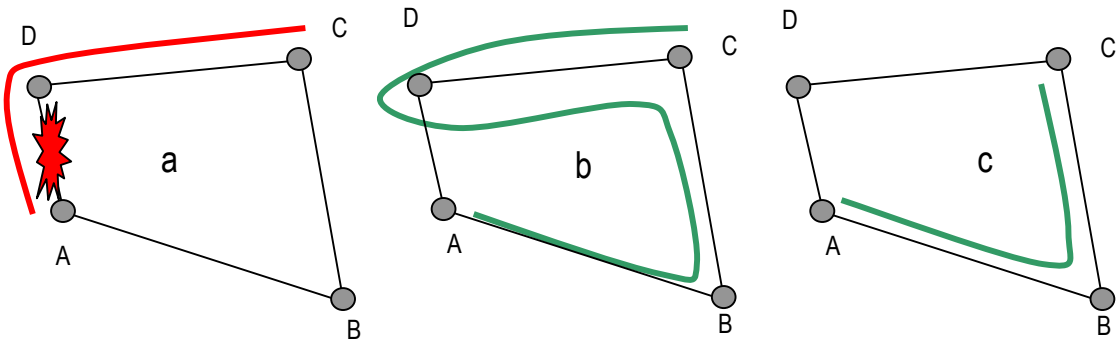


Figure 4.2. 3 ábra . Link (b) and Path (c) Protection for failed link A-D.

Link- or node-disjoint protection?

In practice the majority of the failures affect links (modeled as graph edges). These link failures are typically result of baggers or other digging machines operating around laid cabels or ducts, or failures of lasers in transmitters and transceivers.

The node (modelled as graph vertices) failures happen rarely, these are results of aircondition failures, long supplay outages or disasters, catastrophs.

Accordingly, the protection and working paths of all node-pairs should be either link- or node-disjoint. The node-disjointness is a stricter criteria, than edge-disjointness. Clearly, node disjoint paths require more resources.

Dedicated or Shared Protection?

Employing path protection with disjoint paths, the working path typically uses the shortest possible path since it is steadily beeing used, while the protection path is the next shortest disjoint path, which is typically significantly longer than the working one. Therefore, assuming dedicated disjoint path protection 2-3 times more resources are needed than without any protection.

Two types of *Dedicated Protection* are distinguished. In case of *"1+1"* protection the information is sent along both, working and protection paths simultaneously. The receiving end will decide which signal to use based on quality monitoring. Inspite, in case of *"1:1"* protection the signal is sent along the working path only, and after detecting a failure, both the sending and receiving ends have to be notified, and both will switch to the working path. This is referred to as (APS:

Automatic Protection Switching). *"1+1"* protection is simpler and faster, however, its operation is more expensive.

Due to very low failure probability the expences of Dedicated Protection are unaffordably high. Therefore, assuming low failure probability and fast repair Shared Protection is preferred. In this case is sufficient to allocate one or *M* (where *M<N*) protection facilities for *N* working facilities referred to as *1:N* or *M:N* protection, respectively.

Assuming that working paths sharing common protection resources may not have common parts, a significant amount of resources can be saved without significant availabilty deterioration.

4.2.6. Special Structures for Network Protection

The methods for enhancing availability of networks we presented in last section are strongly topology and traffic dependent. The network density, for example, significantly influences the efficiency of the above mentioned methods. Protection of rings is very simple, fast and based on local decision, i.e., it works in distributed (in contrast to centralised) manner. Therefore, even in mesh networks the operators form logical rings to obtain better protectability. This method is used particularly in SDH/SONET networks. However, the drawback of this method is its significant resource requirement. In mesh networks is preferable to use the thriftier mesh restoration.

Some additional resilience structures worth mentioning are the *metamesh*, where the whole network is simplified by aggregating neighbouring nodes of degree two, or defining protection cycles (*p-cycle*), where the traffic of failed paths is always redirected to them, or by configuring protection paths (*p-tree*), where the network is covered by two trees in such a way, that in case of any single failure all nodes are accessible along at least one of the two trees.

The increasing amount of traffic and the evolution of optical technologies has resulted in building networks consisting of two or more layers as, e.g., IP/ATM/SDH/DWDM or similar structures.

IETF has proposed GMPLS (Generalised MultiProtocol Label Switching), while ITU-T has specified the G.astn-t (Automatic Switched Transport Network). In these

multi-layer networks there arises the question what layer should perform the protection or restoration? The best approach is that always that layer should act which was primarily affected by the failure. However, in this case all layers must have enough spare resources for bypassing failures. All layers above the failing layer will be affected by that failure. Therefore, it is very important to avoid protection in all affected layers.

When *traffic grooming* is used it is advantageous to handle these traffic streams jointly, i.e., protecting the whole groomed traffic. This requires more capacity, however, it is faster and requires less signalling than in case of switching paths one-by-one, i.e., without grooming.

The network resilience concept applied in the network strongly depends on the network dynamicity, i.e., whether it is a statically configured leased-line-type network, or is it a dynamic on-line switched network. While in the first case the resilience management can be centralised, in the latter one not really, since the network state is steadily changing. Furthermore, in case of dynamic networks a huge amount of state information has to be fludded. If shared protection is applied even more state advertisment traffic is beeing generated. Not only the link state and topology are advertised, but also both the working and protection path of each node pair has to be known in the whole routing area. In this case the network has to be partioned into smaller areas or, the information has to be aggregated. Typically Automatic Protection Switching (APS) is used in terminating nodes to bypass the failures by switching the affected traffic to protection resources.

4.2.7. How to estimate availability?

The network consists of network elements. For example, a link connecting two cross-connects or switches consists of regenerator sections, regenerators, receivers, transmitters, and line terminals that can be all considered network elements. The network nodes consist of a matrix, a clock, and some controll and management units.

For all the above listed elements the *availability* can be estimated separately. For example if a 300 km long cable is laid along a railway, statistically it is cut once every seven years. Repairing it takes typically one day. The availability (probability that it will carry oout its mission properly) p_i of this link can be estimated as follows.

Let MTBF denote the Mean Time Between Failures (7 years in our example) and let MTTR denote the MeanTime to Repair (1 day in our example).

Now the availability of element *i* can be determined as

$p_i = 1 - MTTR_i / MTBF_i$,

In our case $p_i = 0.9996$, i.e., the considered link is available in the 99.96% of time.

For simplicity reasons the failure probability can be considered uniformly distributed along the cable, e.g., if the link is only 100 km long instead of 300 then the availability of this link will be 0.9999. As can be seen the time needed for repairing the failed link significantly influences the availability value, therefore, it has to be added as well.

In a network a connection uses numerous network elements along its path. Even if one of these fails, the whole connection will fail as well. Availability of such a system consisting of *I* (serially connected) elements can be evaluated as

$$p(i=1,2,\ldots I)= p_1 p_2 p_3 \ldots p_I$$

It can be seen that the availability of the whole serial system is worse than that of the element with worst availability. It can also be seen that the longer the path is the worse its availability is. The length of the path implicitly means the number of elements it consists of, as well.

How can then availability of "five nines" be achieved? There must be parallel protection facilities which can overtake the traffic of the failed path within miliseconds. Let us assume 1+1 protection is beeing used within the network. Then the protection works in, e.g., 10 ms. This time can be neglected compared to 1 day out of service in case without protection. Furthermore, the connection is considered to fail only, if both the working and protection paths fail.

Let now p_i ($i=1\ldots I$) denote the availability of certain paths. Then the availability of a system consisting of a working and *I-1* protection paths can be formulated as

$$p(i=1,2,\ldots I)= 1 - (1-p_1)(1-p_2)(1-p_3)\ldots(1-p_I)$$

It can be seen now that the availability of the whole system will be better (higher) than the availability of the paths the connection relies on. The more

protection paths are available the higher the availability will be. This allows guaranteeing very strict availability requirements even in networks consisting of often failing elements.

References

[4.2.1] Cinkler T., Henk T., Ziegler G. ed.: Távközlő hálózatok (Telecommunication Networks), electronic lecture notes, BUTE, 2002, http://leda.ttt.bme.hu/~cinkler/TavkHal/

[4.2.2] Hosszú G.: Internetes médiakommunikáció (Internet-based Mediacommunication), *LSI Oktatóközpont,* Budapest, 2001.

[4.2.3] S. Keshav: An Engineering Approach to Computer Networking: ATM Networks, the Internet and the Telephone Network, *Addison-Wesley,* Reading, Massachusetts, US, 1997.

[4.2.4] H.G. Perros: An Introduction to ATM Networks, *John Wiley and Sons*, Chichester, 2001.

[4.2.5] A.S. Tanenbaum: Számítógép-hálózatok (Computer Networks), *Panem-Prentice Hall,* Budapest, 1999.

References for Sections [4.2.5] [4.2.6] [4.2.7]:

1. T. Cinkler, T. Henk, G. Ziegler szerk.: Távközlő hálózatok, elektronikus jegyzet,BME, 2002, http://leda.ttt.bme.hu/~cinkler/TavkHal/ (in hungarian)

2. A.S. Tanenbaum: Computer Networks, Prentice-Hall, 1996

3. A.S. Tanenbaum: Számítógép-hálózatok, *Panem-Prentice Hall,* Budapest, 1999

4. B.S. Dhillon: Reliability in Computer System Design, *Ablex,* 1987

5. A. Kershenbaum: Telecommunications Network Design Algorithms, *McGraw-Hill*, 1993

6. R. Bhandari: Survivable Networks: Algorithms for Diverse Routing, *Kluwer,* 1999

7. D.P. Bertsekas: Network Optimisation: Continuos and Discrete Models, *Athena Scientific*, 1998

Translated by: Krisztián Németh

4.3. Network planning

Tivadar Jakab, author

László Jereb, PhD , reviewer

Planning of telecommunications networks is based on the sophisticated application of complex scientific methods to design and calculate the optimal configuration and amount of network resources according to the techno-economic specification and boundary conditions specified in advance.

In the traditional network planning process quantitative requirements are derived from service forecasts with help of analytical traffic models. The process itself is focused on resource dimensioning, equipment specification and configuration according to the specified qualitative requirements derived from desired service quality.

The planning approaches of the recent practice consider three different time horizons, long-term, medium-term and short term ones.

Long-term plans are focused on the strategic network vision, on the definition of quasi-permanent network characteristics, fixed or slightly changing ones on a longer run, like logical network structure and fundamental network infrastructure. Medium and short-term plans are aimed to specify network developments and investments.

Network plans with different information structure and details correspond to different motivations and scopes of different planning time horizons. It is a key issue whether the approaches, models and methods applied in consecutive processes realise a coherent, economical and effective network operation and development.

Network analysis (focused on throughput, reliability/availability or overall performance) is applied to check and validate the planning models and results, and to study the network behaviour under conditions different from the ideal/nominal assumptions (like e.g. traffic overload, network element failures, etc.).

The definition of an optimal logical structure (configuration) is one of the main goals of the network planning process. However, network engineering, the calculation

of the physical parameters of transmission routes and the definition of interfaces and inter-working conditions for hardware and software network elements are important parts of the planning process, as well.

During the last decade the telecommunications networks have been permanently migrating, and this process is likely to continue in the near future, as well. Each of the new networking technologies and the fundamentally new services, as well as the changed market conditions could result in significant modifications in telecommunications networks, and their accumulated impact is a permanent change.

As a consequence of the permanent migration described above, a long-term co-existence of different network technologies is foreseen. Thus, the key issues of networking in the near future are the elaboration of proper migration strategies, and the realisation and maintenance of optimal co-operation conditions of co-existing technologies.

4.3.1. Changes in planning approach

Due to the changed techno-economical conditions the approaches, processes and methods of logical network planning are changing significantly.

The fast progress in technologies, the shortened innovation cycles, the hardly forecastable environment and the uncertain technical and marketing conditions result in closer planning time horizons and a change in the goals and contents of the network plans. As a significant consequence, the increased importance of strategic planning can be identified. However, strategic network visions only orient a given phase of the development processes, since the reconsideration of the strategic goals and visions could be necessary several times during a long-term planning period.

The new achievements in the development of technologies and services and the changing market conditions may require significant modifications in the main characteristics of networks and service provisioning. These required modifications could be enabled in long lasting network development processes based on accurately developed and maintained strategic network visions. The long-term network vision identifies an optimal target network concept taking into account both current and forecasted requirements and conditions (Figure 4.3.1).

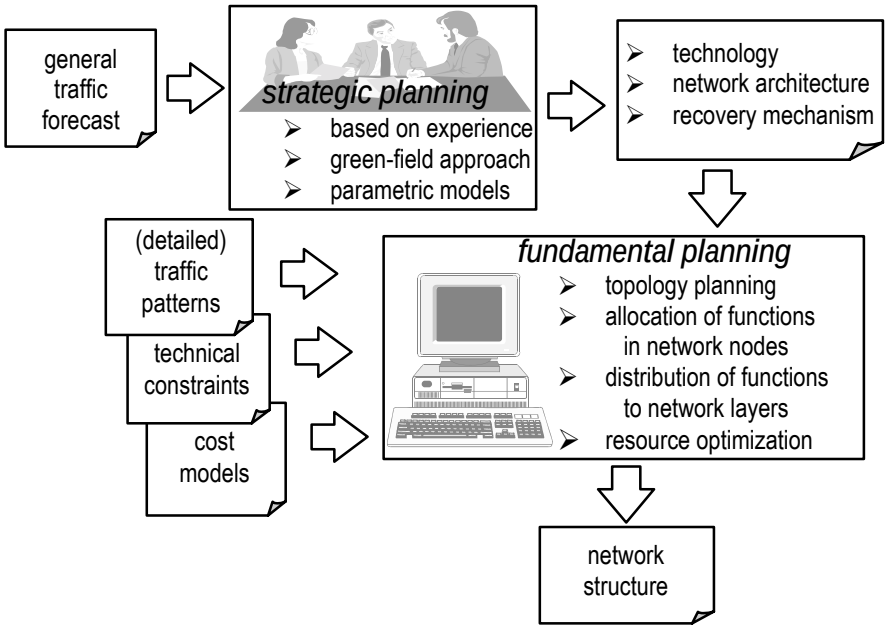
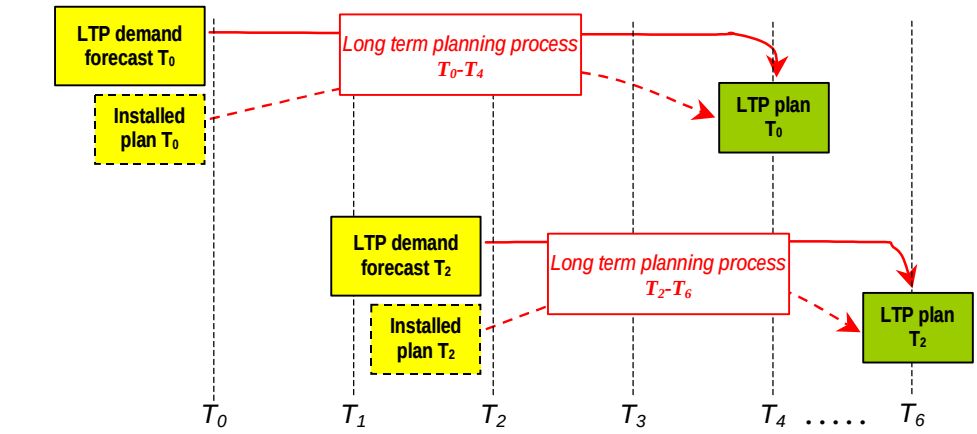


Figure 4.3.1 : *Strategic and fundamental planning*

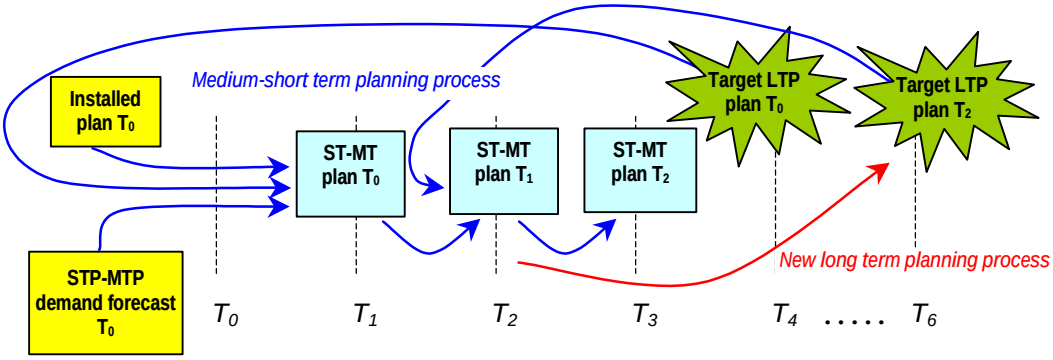
Another significant consequence of the changes in planning time horizons is that the short-term planning considerations are dominating the planning process. During recent years the medium-term planning cycle has become shorter and shorter, and has been practically merged with the short-term one.

The relation between the long-term and medium-term planning processes is depicted in Figure 4.3.2. According to the medium-term planning approach illustrated in Figure 4.3.2.b the target network vision (set up in a process illustrated in Figure 4.3.2.a) is only a driving force, and the medium-term planning cycles are not aimed at the gradual implementation of the target network. Following the illustrated concept, a medium-term plan is only a step towards the network specified by the target network vision. In such a complex planning process the long-term planning is repeated and the target network vision is reconsidered according to the conditions and requirements changed during the long-term planning period. The approach described above is called rolling technique in planning. The main idea of this technique is that before the end of a long-term planning period (T_4 on Figure 4.3.2.a) the target network vision is adapted to the changing conditions and requirements (in T_2). The corresponding medium-term periods overlap with the time window defined by the long-term planning horizon. The motivation for that is given by the significant potential changes in the determinative planning requirements and conditions (technology,

service demands, economical and market conditions) during a long-term planning period.



a) Long-term planning approach



b) Medium-short term planning approach

Figure 4.3.2 Long and medium-short term planning processes

The set of basic problems covered by the planning processes is extended. The emphasis on tactical planning is increasing, since the analysis and evaluation of the existing network infrastructure, resources and configuration are key issues in a permanently changing environment. The role of proper network analysis methods and techniques is fundamental in these tactical planning processes.

There are two options to cope with the permanent changes and the difficulties in traffic (mainly Internet traffic) modelling and forecasts. Either the network should be over-dimensioned significantly to facilitate control over the unforeseen changes, or based on the increased intelligent network flexibility the planning paradigm should be modified according to a measurement based network planning and configuration concept. The basic idea of this concept is to change the configuration of existing network resources in order to improve service quality and network efficiency on the

short run according to the requirements and trends derived from the current traffic measurements. If the service quality and network efficiency expectations cannot be met, the network development needs and requirements can be identified, and in the next network extension (investment) cycle the required resources can be installed.

There are two fundamental requirements to set up the process described above: the fast and automatic configuration of network resources supported by the applied network technology and management, and the amount of some extra resources enabling the adaptation of the network to the changing requirements.

The foreseen long-term co-existence of different technologies enabling different implementation of same or similar networking functions implies changes in network planning, as well. The decisions on the applied technology stack and the optimal assignment of fundamental networking functions (resilience first of all) to specific network layers have become fundamental planning issues. The assignment of different functions to different network layers is influenced significantly by the inter-working capabilities of the applied technologies.

The development of planning methodology and software tools is usually behind the technological progress of networking. Now, therefore it is a fundamental requirement that these methods and tools should be based on unified general models and approaches. To meet this requirement the new methodology should be based on technology-independent models and layered networking approaches.

4.3.2. Consistence of planning processes

To have a successful and effective network development process the coherent mapping of network planning processes targeting different time horizons is fundamental. Since it is not obvious to provide the required coherence in the planning processes of increased complexity an overview of potential incoherences and the measures against them is given.

One of the sources of potential incoherence is the uncertainty of different forecasts. The longer the considered period of time the more uncertain the forecast is. This often results in incoherence of corresponding planning results derived from different forecasts. The main problem originates from the forecasts of service/traffic demands, however, it may arise concerning the forecasts of economic and

ncoherence cause	Between
High uncertainty	LTP - MTP and MTP -STP
Budget limitation	LTP - MTP
Different optimisation criteria	LTP - MTP
Different resource allocation rules	LTP - MTP and MTP -STP

Table 4.3.1. Causes of incoherence between planning phases

technological trends, as well. The unforeseen and unexpected significant changes in service/traffic patterns may cause significant difficulties in network development processes.

Another typical problem resulting in incoherence is the potential lack of financial resources, or the significant changes of the budget assigned to a given network development phase. As a consequence of such problems the installation of new network elements are delayed and the planned (optimal) schedule of network development cannot be kept.

The third kind of problems originates from the different optimisation criteria, approaches, and the differences in the resource allocation strategies applied for different planning periods. The potential impacts originating from these differences are increasing with the sharp competition in the telecommunications market.

The unfavourable consequences of different incoherences depend on the planning processes the problem arisen between (Table 4.3.1).

The incoherence between the long-term and medium-term planning weakens or may even eliminate the driving force of strategic decisions, but does not have

Incoherence cause	*Preventive countermeasure*
High uncertainty	Improving the forecasting methodologies
	Reducing the considered time-scale
Budget limitation	Improving the overall network creation process
	Demand reconfiguration
Different optimisation criteria	-
Different resource allocation rules	Improving the overall network creation process
	Reducing the considered time-scale
	Demand reconfiguration
	Selecting simpler network structures and architectures

Table 4.3.2. Preventive countermeasures against incoherence of network evolution

direct impact on the current market positions.

However, the incoherence between the medium-term and short-term planning may disable the satisfaction of certain capacity demands, and directly influences the current market positions and quality of services.

There are preventive and corrective measures to protect the unfavourable impacts of these planning incoherences summarised in Table 4.3.2.

4.3.3. Measurement based planning

There is a significant set of uncertainty in IP network planning:

- New applications and services without detailed traffic models.
- The traffic modelling of traditional applications and services are complex and sophisticated and the application of the models in the planning and dimensioning processes is quite difficult (e.g. the identification and matching of traffic parameters).
- It is difficult to identify or estimate the interest pattern of the traffic offered at a certain network node.
- Since Internet traffic was rapidly growing in recent years, the forecast of the amount and the growth rate of offered traffic are quite uncertain.

Due to these difficulties the applicability of traditional network planning methodology and network development strategies based on this methodology is an open question. To reduce the significant uncertainty in network development a new planning approach is required. The fundamental idea is to replace the traffic forecasts by the information derived from traffic measurements and network utilisation data. Based on this idea a complex network planning and configuration process can be constructed, where the needs for automatic network configuration or capacity extensions can be derived from the measured network traffic characteristics.

The schedule of interventions is a key issue in measurement based planning processes. Between two interventions the measurement module evaluates the network traffic characteristics. The time between two interventions strongly depends on the dynamic changes of the demand pattern and on the performance of the network.

The proper data processing methodology is essential in the measurement based planning and configuration processes. This methodology is required to derive traffic characteristics from the directly measured data, so that these characteristics could be applied to define the current network condition and the optimal network configuration and to identify network extension needs. The characterisation of network traffic is based on data from active and passive traffic measurements, and the traffic information accumulated by the network elements. Complex mathematical theory and methodology are applied to set up proper traffic models and traffic estimations.

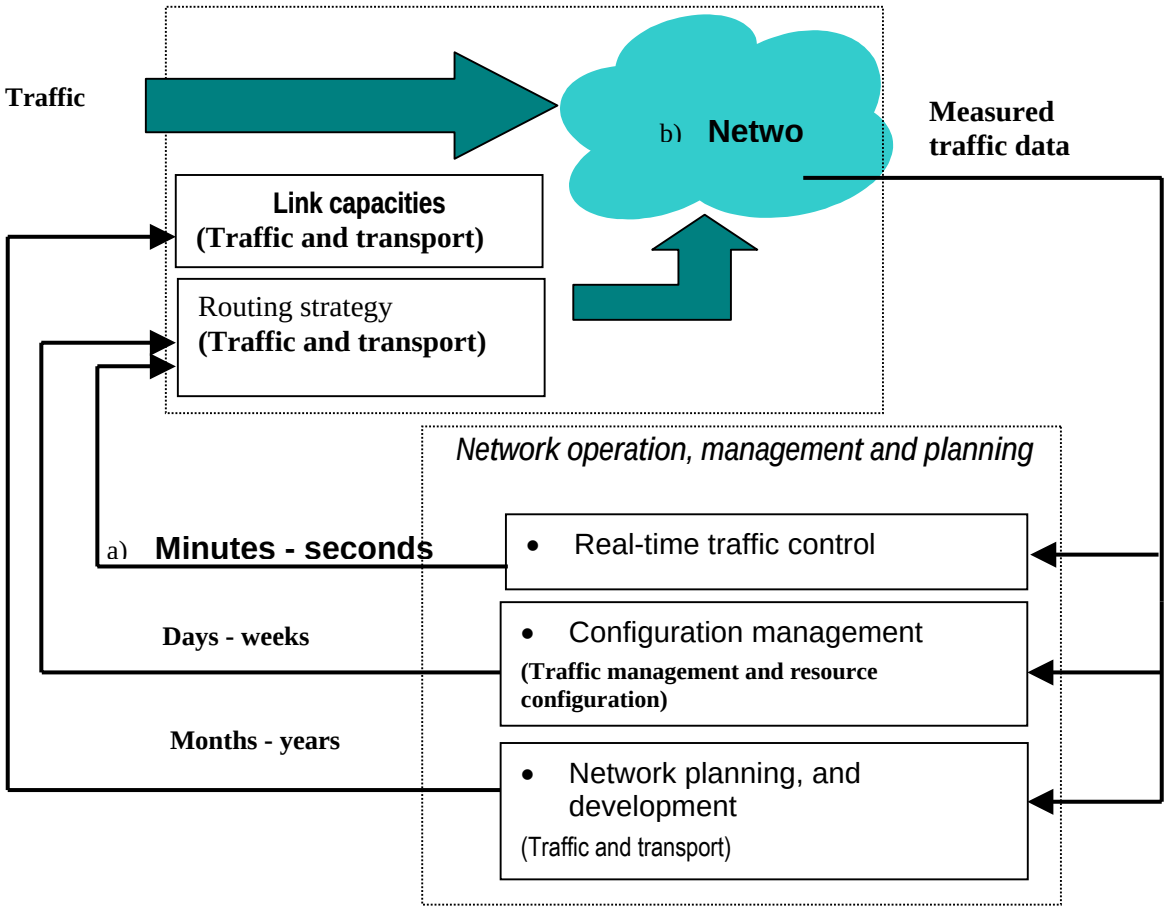


Figure 4.3.3 Complex scheme of measurement based network management, configuration and planning

Figure 4.3.3 depicts a measurement based complex system of network management, configuration and planning.

4.3.4. Planning problems and methods

Although the pure technical content of the telecommunication network planning itself implies a complex and sophisticated process, there are numerous further requirements and restrictions to be considered impacting the formalisations and solutions of the planning problems significantly. (Only a short overview of the issue is given here, for more details see reference [4.3.3].)

General features of network planning problems

Three sets of aspects are identified according to their impacts on the characteristics and content of the planning problems:

A) *The permanent progress of technologies*: The technological advance and change facilitating different networking solutions have significant impact on network planning. Let us take two simple examples for illustration. Technological changes on the switching network layer from circuit switched technology (PSTN - Public Switched Telephone Network) to cell switched technology (ATM - Asynchronous Transfer Mode), then to packed switched technology (IP) results in fundamental changes in networking. These changes have direct impact on network planning, as well (e.g. concerning the applicable traffic and queuing models). The progress of wired technologies providing permanent transport capacities (PDH - Plesiochronous Digital Hierarchy, SDH - Synchronous Digital Hierarchy, WDM - Wavelength Division Multiplexing) enables new network functions and structures to be realised. Let us just remember the changes of preferred network topologies - meshed, ring based, again meshed – to identify the impact of technological changes on network planning. On the one hand these permanent changes require the renewal of planning models and methods to enable the investigation of the capabilities of new technologies in an early stage. Based on these studies the impact of the new technology on the strategic network vision can be identified. The other significant consequence of the permanent technological progress and rapid changes is – as it has been described earlier – the long-term co-existence of different technologies in the network. It can be difficult to describe, model and solve numerous planning problems originate from the technological co-existence.

B) *The current economic environment and market regulations*: The direct impacts of the economic environment can be identified on the financial background, on the conditions of economical operation, and on the risks concerning the longer-run return of the investments. However, via these impacts the economic environment may affect the technical solutions, as well. For example, in case of the appearance of a new technology based market segment, the introduction of this new technology without

providing any real short-term economical and technical advantages may become a strategic issue just to get market shares from this new segment.

C) *The characteristics of the organisation that intends to use/operate the network* to be planned may have significant impact on the planning. The cogency of permanent renewal, the innovative capabilities of the organisation and the flexibility to accept new technologies are important aspects to be taken into account in an extended planning process. A characteristic example illustrating the importance of these issues is the controversial reception of IP technology. The experts from telecommunications fields and the experts from computer science and computer networking fields have significantly different attitudes to the conquest of IP technology. The strength of the technology orientation and the cost sensitivity of the given organisation has a significant impact on the desired networking solutions. The dominant user or service provider attitude of the organisation approaches (a private network - the users' investments according to their own needs or a public network - to provide services under market conditions) results in fundamentally different planning approaches. However, the methodology applied to plan private or public networks is derived from a common basis.

A set of characteristics having a decisive impact in network planning processes can be identified independently from the different technologies, the different applications and the different networking strategies and solutions:

- The stochastic behaviour of service demands together with the server capabilities and processes of the applied technology determines the queuing models applicable for the given network.
- The telecommunication environment with many service providers, with sharp competition and with networks of different aims, technologies and structures has an impact on the planning through technical, service oriented and operational co-operations.
- The growth trends identified and forecasted imply decisive planning aspects (e.g. extendibility, scalability, efficiency, and return of the investments).
- The wide choices of hardware and software products that can be applied in the implementation of different networks arise significant planning problems concerning modelling and planning methodology and co-operational aspects, as well.
- The capabilities facilitating and the restrictions limiting the co-operation of different network elements in the networks both inside one operator's domain or across the domains of different operators, determines the potential networking solutions.
- The co-operation related recommendations specifying the optional technical solutions can be applied for different networking problems.
- The risks originate from the long run return of huge telecommunications investments (flexible upgradability, economical operability) have significant impacts on the economic considerations of the network planning.

Optimisation criteria and tasks

There are two main general approaches to network planning problems:

- One of them is focused on dimensioning, configuration and installation of network resources of economical size to achieve clearly specified service quality requirements. In case of deterministic transport capacity demands (semi-permanent connection based services) line systems of economical sizes and filling rates can be realised applying different multiplexing techniques (time, wavelength, etc.). In case of random traffic demands (switching based services) the main objective is to achieve satisfactory utilisation of network resources, which can be realised with help of different statistical multiplexing schemes. These kinds of planning approaches are aimed to meet a set of qualitative (demands) and quantitative (quality of service) requirements with minimal amount of resources.
- The second one covers the case of limited network resources, when the objective of the planning is to maximise the network performance. The potential solutions are restricted by the available network capacities (the existing infrastructure) or by the limited financial resources dedicated to the problem. The objective of the planning is to maximise the network efficiency in both variants.

  The solutions of typical planning problems depend on:

- the required planning details,
- the main characteristics of the network, the supported services and the applied technology, and
- the time horizon of the planning.

  There are four sets of typical planning problems:

The classical planning problems belong to the first set (to meet given specifications with minimal costs). These problems can be present in planning as directly specified tasks, or they can be embedded in complex problems, like for example the optimisation of logical or physical network structures.

In a different set there are the network configuration and reconfiguration problems. As for these, the objective of the planning is to maximise the network performance (e.g. the total number of served users, total throughput, availability, etc.) using the limited resources of the existing network.

The third set comprises problems that can be considered as the combinations of those from the two previous sets. These problems are focused on the optimal network extension, and can be solved by the optimal utilisation of existing resources and the optimal allocation of new ones. These solutions can be restricted to the

utilisation of the free existing resources only, or can be extended for the reconfiguration of the entire network.

In the fourth set there are complex problems. The evaluation of the candidate solutions of these problems requires the dimensioning problem to be solved. These complex planning problems are focused on the optimisation of the network structure. During the optimisation of logical and physical network structure optimal location problems (concentrator, switch, etc.) and topology optimisation problems (duct, cable and link topologies) are to be solved.

4.3.5. A general layered model for network planning

To solve different planning and analysis problems of multi-technological networks a uniform network model is desirable. A uniform model enables coherent approaches to different problems and supports the applicability (reuse) of planning and analysis methodology (and the supporting software tools) for networks with different technologies.

A potential technology-independent network model, planning and analysis approach can be based on the functional similarities of the different technologies and the structural similarities of telecommunication networks. The stochastic traffic

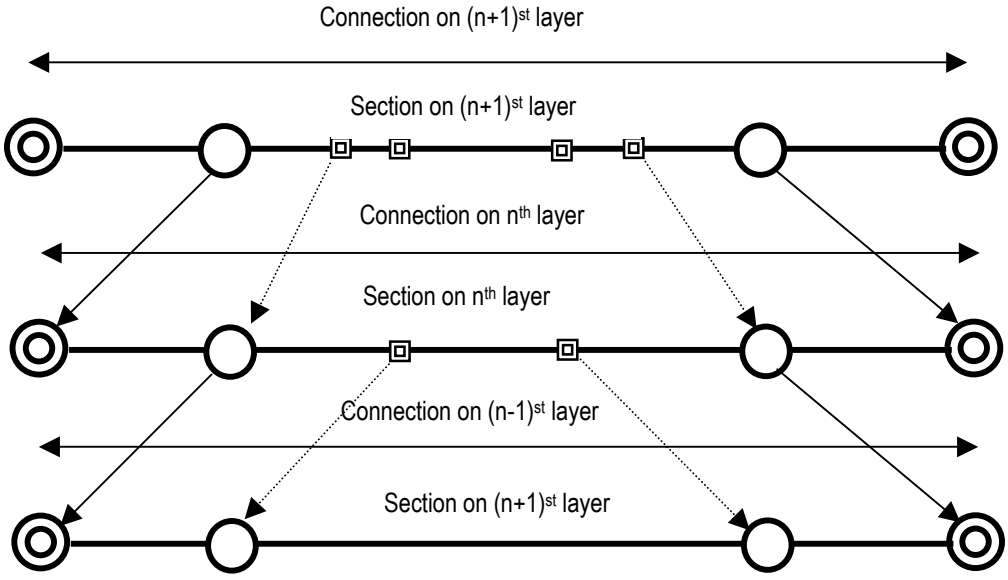


Figure 4.3.4. The simplified layered model of the technology-independent planning approach

demands are concentrated by switches, and the deterministic capacity demands are multiplexed into bundles.

In case of stochastic traffic demands the aim of the concentration is to achieve the best utilisation of shared resources (under specified quality of service conditions) maximising the statistical gains. The multiplexing of deterministic capacity demands results in better filling rates of large capacity modular network elements (multiplexed line systems) improving the scale of economy related gains.

The technology-independent modelling approach is based on the generalisation of the layered network model (G.803, G.805, and M1400). According to this approach, in a hierarchical layered structure each layer forms a layer-network, which relies on the resources of the underlying layer of higher level concentration or multiplexing. The relation between the neighbouring layers is defined by the connections (sections) of the higher (in logical meaning) layer bundled (multiplexed) into connections on the lower one. The conditions for the bundling (multiplexing) are defined by the functions realised in the given technology and installed at different network nodes. The main characteristics of the applied technology like for example modularity of equipment and operational and economic considerations may have a significant impact on the solutions, as well.

For example, in case of wavelength division multiplexed networks with permanent connections the optical channels can be bundled if the network nodes are equipped with multiplexing, demultiplexing, and with wavelength conversion functions (if needed). The size of the multiplexed bundle depends on the sum of the multiplexed channels.

In case of line or packet (or cell) switched networks proper switching functions are required to perform statistical multiplexing. The size of the concentrated bundle can be calculated based on the technology specific traffic models and formulas derived from the queuing theory.

Planning multi-layer networks

When operating multi-technology networks it is fundamental to enable optimal inter-working environment for the different technological layers, thus, the inter-working capabilities are key issues in the planning process, as well.

The inter-working of technological layers is based on co-ordination. The different technological layers can be operated independently, without any co-ordination, but it may result in redundancy concerning resource allocation and instability in operation (e.g. dispreferred redundancy of resilience related resources, malfunction of countermeasures in case of failures).

The simplest solution to implement co-ordination is timing. In case of timing based co-ordination the layer with longer reaction time waits for the faster one, and starts reacting after a certain delay (of a length depending on failure cases).

A stricter co-ordination can be achieved through changing information between the involved layers (token-based co-ordination), however, the optimal co-ordination requires integrated operation and management systems. It is difficult to integrate the operating and management systems of network layers with different technologies if the interfaces and processes are not standardised in details. The integration may require specific hardware and software developments, and may result in difficulties concerning scalability and may strictly limit the set of potential manufacturers. The progress in telecommunication technologies approaching an intelligent network with unified signalling – the GMPLS (Generalised Multi-Protocol Label Switching) based ASTN (Automatic Switched Transport Network) – tries to solve the inter-working problems.

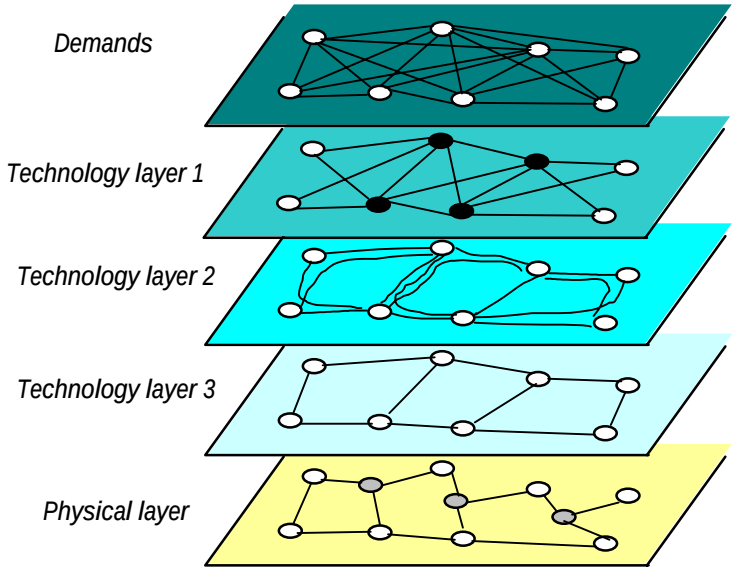


Figure 4.3.5. Layered structure of a multi-technology network

Networks comprising different technological layers arise fundamentally new planning problems. The most critical ones of them are how to assign optimally the network functions between the layers in client-server relation, and how to build a network structure with optimally co-operating layers.

A top-down planning approach fits well the layered network model based on the generalised multiplexing scheme, where the requirements of a server layer are expressed with help of point to point demands and their realisation rules. This traditional approach enables to take into account the planning features of each technological layer optimally. However, it is quite difficult to plan the implementation of similar network functions in different layers, for example, planning the resources for protection and restoration according to this approach may result in dispreferred redundancies.

A different approach is the bottom-up technique. It seems to be a heresy to plan and dimension a server layer without clearly specified client demands and requirements. Can it be considered as a realistic alternative?

Developing a network plan this way, the lack of information on clearly specified demands and requirements may restrict the elaboration of an optimal solution. However, the relatively low cost and high capacity optical transport technology may enable a change in planning paradigm, increasing the emphasis on functional and structural planning instead of the capacity optimisation.

The third potential strategy is the joint planning of different layers. Through applying this strategy a more optimal solution can be achieved than those of previous strategies, however, there is a certain risk that the parallel consideration of client layers with different planning features and requirements requires a very sophisticated and complex planning process.

Based on the above considerations it seems to be an expedient compromise to apply an iterative approach in planning, which enables an improved and effective process both in top down and bottom up cases.

The fundamental changes in telecom markets, the permanent advances in technologies and services require the reconsideration of planning approaches and methodology.

The shorter innovation cycles, the difficulties in modelling and forecasting result in increased importance of analysis methods, and put tactical planning and measurement based planning in the foreground. Besides the short term planning approaches regularly adapted and improved strategic plans keep their importance in long-term orientation of network development processes.

Taking into account the relatively cheep optical technology based transport it seems to be a promising option to take an approach different from the traditional capacity optimisation. This approach results in a planning process, which targets to define a flexible network structure able to meet varying client demands and requirements, instead of a network with optimal capacities derived from well-defined demand forecasts. The strategy outlined above enables the planning of a server layer, which is operated according to the scheme described as a measurement based network configuration and planning. As far as the driving tendencies can be identified in the advances of technologies the IP client and the automatically switched optical networks (ASON - Automatic Switched Optical Network) will realise network architecture close to the concept outlined above.

4.3.6. Network optimisation methods

Since the problems of network planning are complex and diverse, the methodology of network planning consists of the integration of different disciplines.

Models and methods of queuing theory are applied to describe the characteristics of stochastic traffic; while statistical methods are used to process the huge amount of network data (e.g. from traffic measurements).

Methods from operations research, mathematical programming and graph theory play important role in the specifications and solutions of optimisation problems. Despite the fact that the applicability of exact methods to solve problems of practical size is limited, these methods provide important references to evaluate more scalable heuristic solutions.

To analyse and design switched networks (carrying random traffic) exact analytical formulas and approximations are derived from queuing theory. When the size and/or the complexity of the problems to be studied are out of the analytical

treatability simulation models and methods can be applied. The overview of the analytical traffic models and methodology are out of scope of the present summary.

The methods to solve planning and dimensioning problems of permanent connection based networks (supporting capacity demands) strongly depend on the applied technology, however, the problems of similar characteristics can be solved with help of the methods of operations research. The exact solutions of basic sub-problems (like for example search for a minimal route, for a minimal spanning tree, for a minimal cut, etc.) are based on graph algorithms, or can be calculated by the methods of integer linear programming (ILP).

The ILP based methods are not effective enough in practice, i. e. to solve problems of realistic size, since they require significant computational resources. However, ILP may ease and support the implementation of solutions for new planning problems. ILP based methods help to develop exact specifications of the problems to be solved, and provide exact solutions as references for the development of approximate methods and algorithms.

Methods to find optimal solutions for different network planning problems can be based on the exact and heuristic methods of operations research.

The technical, service and economical requirements and economic conditions specify the set of allowed solutions for a given planning problem, and the target function derived form the planning problem gives the criterion to find the optimal one among them. To solve a network planning problem the evaluation of a part or the entire set of the allowed solutions according to a given strategy is required.

An obvious solution is the exhaustive search. Each one of the allowed solutions is systematically visited and evaluated to find the global optimum. The method is simple, but the huge number of potential solutions strictly limits the application in realistic cases.

Having found an appropriate modelling and problem formalisation efficient estimations can be given for the target function enabling the gradual reduction of the subsets of solutions containing global optima. Based on efficient estimations significant amount of allowed solutions far from the optimum could be excluded without evaluating them directly. This optimisation strategy is called *Branch and bound*, referring to the gradual extension of the solutions involved in the evaluation

and to the estimation based decrease of the number of solutions kept in scope. The method is more efficient than the exhaustive search; however, sometimes it is difficult (or requires significant simplifications) to find proper models and to construct effective estimations for certain planning problems.

Instead of visiting the complete state space of the allowed solutions, strategies can be based on search along a given trajectory (specified by heuristic considerations). Adjacency of solutions should be identified (which solutions are considered neighbouring, and how to derive one from the other), and starting form a given solution each neighbouring one should be derived and evaluated to define the direction of the next step via the trajectory.

Three different realisation of the strategy is outlined. On of them is the greedy algorithm, which is based on the evaluation of each neighbour of a given solution, and decides on the most improving one. The method can be easily implemented, it is fast, however it achieves local optima.

The Taboo search realises a systematic search, as well, based on the evaluation of each neighbour of a given solution, and takes the best one according to the target function, however the crank-back is prohibited (taboo). The provided solution is the best one from the visited solutions.

The gradient method is a third method for systematic search, which requires a differentiable target function visiting solutions according to the gradient direction.

Further heuristics based on random search strategies are available. The classical Monte-Carlo method realises a pure random selection of the states in the space of allowed solutions, providing the best from the visited ones.

The hill-climbing method is based on the evaluation of a randomly chosen neighbouring solution, which is taken as the next actual one, if the target function value belonging to the neighbour is better.

The simulated annealing evaluates the target function belonging to a randomly chosen neighbouring solution, as well. A solution with better target function value is always taken; however, there is a finite probability to take a neighbouring solution with worst target function value. To move towards a worst solution enables to escape from local optima. The convergence speed of the process and the distance of the found sub-optimum from the global one are tuneable with help of parameters.

It is quite easy to construct a simulated annealing algorithm for a certain optimisation problem (an initial solution is needed, adjacency of the allowed solution should be identified, the evaluation of the target function is required). However to set the optimal parameters for the given specific algorithm may require long trials and complex studies.

Genetic algorithm is based on the analogy of natural evolution. It realises the transmission of the useful and good characteristics from one solution to another (selection, crossover, mutation in a given population). Or, if it is difficult to implement the transmission of good characteristics, the method is limited for pure random search (mutation only), but based on solutions with good characteristics in both cases.

It is practically impossible to solve realistic network planning problems without effective software tools. These tools incorporate advanced database techniques to process integrated network description information for network operation, management and planning. Advanced object oriented and component-based software technologies are extensively applied to implement and develop network planning and analysis software tools.

References

[4.3.1] ITU Recommendations: G.ASON, G.ASTN, G.803, and G.805

[4.3.2] EURESCOM P709 D3: Planning Optical Networks: www.eurescom.de

[4.3.3] Robertazzi, Thomas G.: Planning Telecommunication Networks, IEEE Press, New York, US, 1999

[4.3.4] Wu, Tsong-Ho: Fiber Network Service Survivability, Artech House, Boston, US, 1992

[4.3.5] Nemhauser, G. L. , Rinnooy Kan, A. H. G. , Todd, M J. : Optimisation, Elsevier, New York, US, 1989

[4.3.6] M. S. Bazara, J.J. Jarvis, H.D. Sherali: Linear Programming and Network Flows, Wiley, New York, 1990

[4.3.7] R. E. Bellman, S. E. Dreyfus: Applied Dynamic Programming, Princeton University Press, Princeton NJ, 1962

[4.3.8] L. Cooper, D. Steinberg: Methods and Application of Linear Programming, Saunders, Philadelphia, 1974

[4.3.9] L. Davis: Handbook of Genetic Algorithms, Van Nostrand Reinhold, New York, 1991.

[4.3.10] A. Kaufmann, A. Henry-Labordère: Integer and Mixed Programming: Theory and Applications. Academic Press, New York, 1977.

[4.3.11] A. Kershenbaum: Telecommunication Network Design Algorithms, McGraw-Hill, New York, 1993.

[4.3.12] R.M.H. Otten, L. P. P. P. van Ginneken: The Annealing Algorithm. Kluiwer Academic Publishers, Boston, 1989.

[4.3.13] L. Kleinrock: Queuing Systems, Vol. I: Theory. Wiley, New York, 1975.

[4.3.14] T. G. Robertazzi: Computer Networks and Systems: Queuing Theory and Performance Evaluation. Springer-Verlag, New York, 1994.

[4.3.15] R. L. Sharma: Network Topology Optimization: The Art and Science of Network Design, Van Nostrand Reinhold, New York, 1990

[4.3.16] D. L. Spohn: Data Network Design. McGraw-Hill, New York, 1993.

4.4. Telephone networks

Zoltán Dely, author

Gyula Sallai dr., reviewer

During more than 125-year history of telephone technique, telephone networks have become the most extended ones in the world. Through a long development of several grades, those met their original purpose, the voice transmission on higher and higher level, adding a continuously extending set of comfortable supplementary services. The network, in the beginning serving voice transmission by its each element, went through a differentiation, the switched traffic network and the transmission network have been separated. The latter became the common transport means for traffic networks of different purposes: data networks, mobile networks, etc. Digitalisation of transmission and switching made the network capable to carry data beside voice, and narrow-band ISDN was born. ISDN may be considered as a multi-service extension of digital telephone network. However, *circuit-switching* the fundamental technique being kept so far has become the barrier for economical realisation of services demanding increased and very changing transmission speed on a common network. The future multi-service integrated network – according to our present knowledge and expectation – will be based on the principle of *packet-switching*.

Mobile networks have taken the leading role in the field of voice services by an extensive development that was accelerated in the recent years. The PSTN and ISDN, though its services may be accessed on fixed end-points, today practically reaches into any building – in case of request. Based on this, telephone network is used as *switched access* solution for networks of other service purpose, and not having own access network that reaches the wide user public. For example, this is also the general means to offer Internet to residential users.

The extended wireline access network is the most expensive segment of telephone network, and consequently it also has a rather high value. By DSL (digital subscriber line) digital transmission solutions, traditional copper-line subscriber

network becomes a multi-usage one: beside telephone or ISDN usage, independently on that, it can be used as a broadband access, e.g. for data networks.

In this sub-chapter telephone network is discussed as a stand-alone traffic network, it is delimited from transmission network, however, their relation is presented. We overview the most important characteristics of telephone network, the circuit-switching, the requirements and factors determining the network. We present the traffic structures and traffic routing systems focusing on solutions used today. Detailing the development history is not our intention, earlier stages are shortly mentioned only when it helps to understand connections.

4.4.1. Characteristics of telephone network

Traffic and transmission networks

Telephone network is a traffic network providing voice transmission tele-services, it consists of exchanges performing switching and circuit groups connecting them to each other.

The *traffic circuit* is capable to carry one call at a given time, it realises two-way communication and its concept is similar to that of a graph-edge: the length and path of it are irrelevant, only that matters which two of the exchanges are connected by it. Telephone network uses the transmission capabilities of the *transmission network* for realising traffic circuits. Transmission network consists of transmission nodes and physical circuits. A transmission node may be a section end-point, where transmission terminal equipment, multiplexes are, or a branching point, where physical circuits of different directions may be connected and branched off by multiplexes and/or cross-connects. The path and – depending on the purpose of discussion – the technology and several other parameters of a physical circuit are relevant. For traffic circuits these physical characteristics are irrelevant as long as transmission channels matching to the interfaces of exchanges are provided with given properties (e.g. 64 kbps transmission speed in digital networks). The relation and different characteristics of the two kind of networks are illustrated in Figure 4.4.1. on a simple 3-exchange example. Figure a) shows the traffic network, and Figs b), c) and d) show three different transmission network realisations for this traffic network.

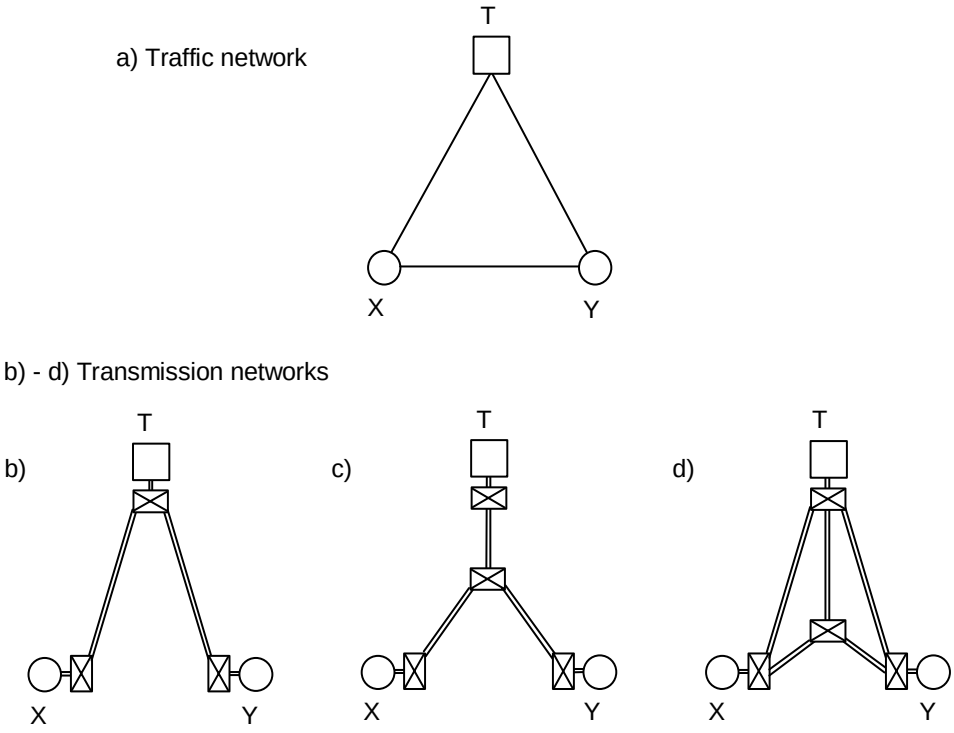


Figure 4.4.1.: Relation of traffic and transmission network

In case of subscriber or *access network* between customers terminal equipment and the subscriber exchange (edge node) the separation of traffic and transmission levels is not common, however with deployment of systems providing multiplex, shared usage, with building multi-service access networks it is in progress (see DSL systems).

Circuit-switching

Making use of telephone network is done by initiating calls. For serving the call the exchanges of the network set up a call-route between the caller and the called party by interconnecting traffic circuits. This route of a given transmission capacity is allocated to the given call for the whole duration of that, independently on the actual transferred information, on the actual intensity of the information changing in time (real circuit based connection, see 4.1.1. section). It shall be emphasised that circuit-switching is executed in the exchanges and it is independent on transmission network technology. Switching of metallic contacts used in the beginning was followed in the digitalisation by matching selected time-slots of incoming and outgoing time division signal-streams to each other. Most generally, in the digital transmission network also time-slots of time division signal-stream represent a circuit,

however, transmission is possible in the form of packet-based (e.g. ATM cells) virtual circuits as well.

Traffic

Telecommunication traffic means the average occupation, usage of a given element, serving unit (switching unit, transmission section, etc.) of the network. In a telephone network traffic is related to the calls, the sum of those, since serving units in the nodes and traffic circuits connecting them are seized when setting up a call-route. It shall be noted that the telephone network occupies circuit groups dimensioned for the forecasted usage, or rather the transmission route of related capacity between each pair of exchanges, permanently. From the point-of-view of transmission network this permanent seizing may be considered as traffic, however, in the discussion of telephone network traffic means the average usage from calls initiated by users, only.

Serving with loss

It is an important feature of the circuit-switched telephone network that it serves the traffic with loss: when a call is stopped on the sections, on the serving units of the traffic routing system available for it, there is no free channel for the next section of the traffic route to be set up then the network refuses the call, that is the call attempt is lost. In other words there is no way for queuing in the serving units of the telephone network.

Factors impacting on the network architecture

When determining the network architecture the fundamental goal is to meet the demands by harmonising technical compliance and economy, that is: to satisfy subscriber and traffic demands according to quality requirements and using technical capabilities in the most economic way. Factors of determining the technically and economically optimal network architecture: subscriber and traffic demands, quality of service requirements, characteristics and capabilities of network elements, and cost parameters of them, given features of the existing network, environmental conditions, implementation constraints and also technical and regulatory requirements, prescriptions.

a) Subscriber demands

- geographic distribution of potential users, future subscribers of the network: subscriber demands are to be forecasted in different market segments, user categories based on their telecom behaviour, intensity and other characteristics of their traffic demand, e.g. distribution on directions;
- distribution of users may be described by continuous function in metropolitan area, and there are areas where demands are located sparsely or grouped in nodes;

b) Traffic demands

- the significant value of traffic demands changing in time that is taken into account for the dimensioning of the network is the *busy hour traffic*, being determined from the daily and weekly process of traffic;
- *specific traffic* parameters for each subscriber categories, that is the average traffic demand per subscriber, and the distribution of it into local, national and international directions;

c) Quality of service requirements

- a part of requirements related to telephone network refer to *voice transmission quality*: noise, attenuation, distortion, delay, etc.
  These transmission parameters and requirements for their value are rather important in analogue networks, they have an impact on the whole network. In the inter-exchange segment of digital networks, meeting these requirements is solved in the transmission network, and structure of the telephone network may be defined independently on the transmission quality parameters. In the subscriber network, in the copper line segments of that the distance limits that can be conveyed shall be considered;
- in the telephone network, the *grade of service* is given by the permitted loss of busy hour calls: what percentage of calls may be refused by the network because of occupancy of serving units;

d) Exchange characteristics and capabilities

- call handling and traffic routing capabilities of exchanges determines the services of telephone network and they have a significant impact on the architecture of the network, as well;
- interfaces of exchanges connecting them into the network, signalling systems and network management systems, capabilities;
- exchanges are complex systems, architecture of them, capacity and variability of different functional modules, and also the maximum realisable capacity have decisive significance;
- circuits of exchanges are parts of call routes being set up, thus transmission characteristics may count (in a telephone network built of digital exchanges and built on digital transmission network the transmission characteristics has no influence on network architecture any more);

e) Cost structure of network elements

- the cost structure of network elements, especially the ratio of cost of equipment in nodes and the distance dependent cost of transmission route has a decisive impact on network architecture.

4.4.2. Traffic structures

Telephone network may be divided into two segments, network parts of fundamentally different characteristics: the *subscriber network* connecting the subscriber end points to the first local (or subscriber) exchange and the *inter-exchange network* consisting of connections among the local exchanges.

Subscriber (access) network

From *traffic aspects* the subscriber or access network has traditionally star-structure: there is one traffic circuit for any one of the subscriber end-points to the subscriber stage, without path-selection. Versions of technical realisation may be found from the copper-line through optical fibre to wireless systems, including combinations of those, as well. In copper-line networks one pair is available for a subscriber. Cost-effectiveness of subscriber network may be improved by applying shared usage elements. Considering specific traffic demands and behaviour of subscribers in a given sub-area, the required grade of service may be reached in certain sections with less number of channels than the number of subscribers by using *traffic concentrating devices*.

Application of electronic devices made possible conveying greater distances economically and thus serving greater areas by one exchange. Important representative of electronic devices performing traffic concentration is the *remote subscriber stage*. Digital switching technology of today is capable to serve rather extended area, even covering several *numbering district* by only one *host* exchange and a number of remote subscriber stages controlled by it.

Inter-exchange network

After determining service areas of subscriber exchanges we have the basic points of inter-exchange network, the local exchanges. Generally, traffic interest is stronger between exchanges within areas defined by administrative, economic, cultural aspects, and it decreases in the function of distance. The economical architecture of the network is a multi-level tree structure, which is matched to a multi-

level area system: a group of local networks constitute a primary network covering a primary area, several primary networks constitute a secondary one, and so on.

In each area an assigned exchange –the traffic centre of the area – *transits* the traffic of the other exchanges in that area, and as a *gateway* it transfers the outgoing and incoming traffic of the given area to and from the others. In this way a hierarchical structure is constituted, and it is the traditional, general structure of telephone networks. For any two of the exchanges one determined traffic route is given, in other words, this is a network of *single connectivity*. Gateways of the areas concentrate the traffic between of distant areas, providing the effective utilisation of long distance network sections.

Along the hierarchy, telephone network is divided into local and interurban network, and the latter is divided into primary district and long-distance transit network. The domestic (national) network is connected to the international network via *international gateway exchanges*. The number of network levels has been gradually reduced as traffic was growing, network elements were developed and the cost structure was changing. The structure of the national network was changed from 5 to 4 levels when introducing crossbar switching in 1974, then it was changed to 3 levels when deploying digital switching in 1990. In both cases the modification was gradual, taking some years.

If in a multi-exchange network or network part there is strong traffic interest between the exchanges per pairs, and there is no traffic-attracting central point, the proper structure is fully *meshed*. In the meshed traffic structure direct circuit group connects any two of the exchanges. This is the effective solution in metropolitan local networks and on the top level of large hierarchical networks.

Structure of the Hungarian telephone network

The national telephone and ISDN network[2] is divided into primary district networks and the long-distance transit network interconnecting the primary networks. The long-distance transit network has two levels, exchanges on the upper, secondary level are fully interconnected (meshed network part). International gateways belong to the secondary level in the network hierarchy. Primary networks has also two levels

[2] The existing status at the beginning of the competition started by the new regulation on 23. 12. 2001, according to the KHVM decree 26/1993.

in principle, however, in most of the districts there are one single exchange. Actual trends – stagnancy or decrease of demands, growing of host exchange capacity – may lead to the further simplification of the network by reducing the number of levels from the existing 3 to 2. In the most cases, switching functions of different hierarchical levels are implemented in one combined exchange.

The local network of Budapest is a special primary network: inner, local transit is made by two tandem exchanges of identical role, gateway function towards the long-distance transit network is realised in separate primary exchanges. Dual transit exchanges improve network invulnerability, provide double connectivity in the considered network. This solution may be economically applied in the transit function of an area when traffic demand is great enough for utilising two exchanges. In the existing network dual transit exchanges are the tandem exchanges and the combined primary-secondary exchanges of Budapest, and the international gateways. Double connectivity may be generalised in the network by connecting each one of primary exchanges to two secondary exchanges, sharing the traffic on them.

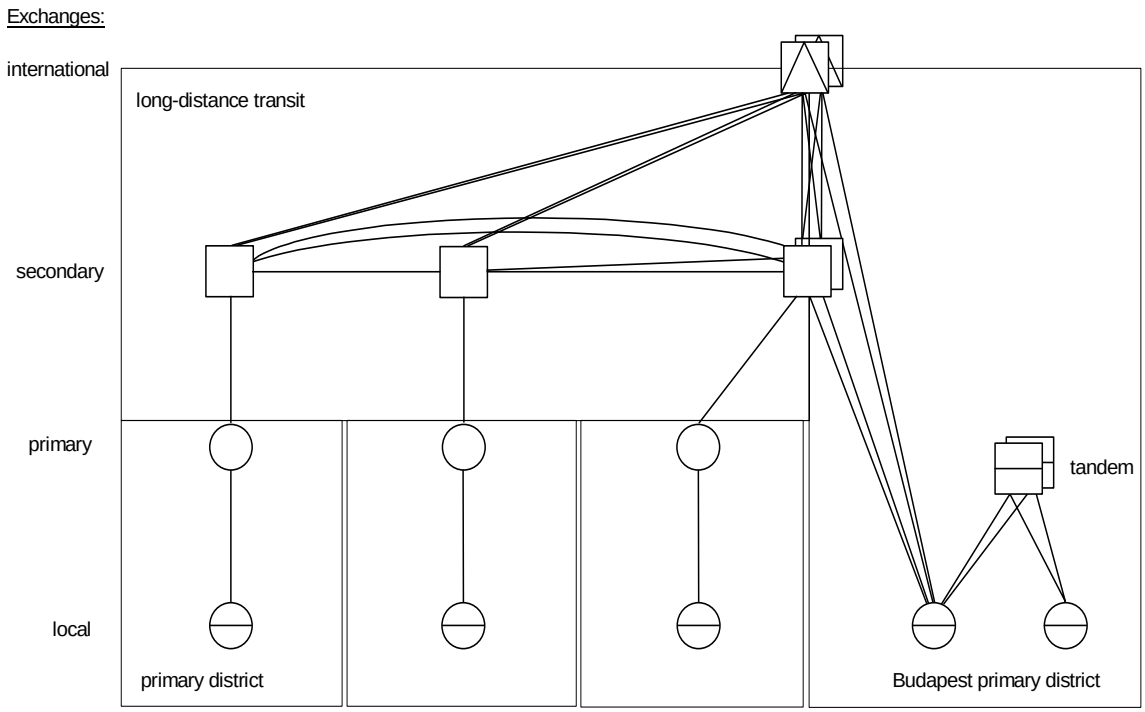


Figure 4.4.2: Structure of the Hungarian telephone network

It is to be noted that, while in the concession period the networks of different service providers matched into the structure described above, those will appear parallel to each other in given geographic areas according to the liberalised regulation. A network partially covering some districts can be built with significantly fewer switching units, than in the present structure.

4.4.3. Traffic routing systems

In the circuit switched telephone network the task of traffic routing is to select the traffic route of calls between the originating and the terminating point. Obviously, it is a task when the network structure offers route selection possibility i.e. generally there is more than one route between exchange pairs.

It was mentioned in the previous section regarding the multi-level tree structure that it gives one single route between any two of the exchanges. This route provides the concentration of traffic belonging to different relations onto common sections, common devices, and thus the efficient utilisation of devices. On the other hand however, for the most of exchange pairs the connection is provided through one or more transit exchanges, which increases the cost of call-routes. If the network is accomplished by so called optional (transversal) circuit groups connecting exchanges, which are not related directly in the hierarchy, it becomes suitable to implement traffic routing aiming primarily the cost-effective construction of the network: in relations having sufficiently high traffic calls are set up without using intermediate switching devices, while the common routes are taken for relations of low traffic.

Call set-up principles

In telephone networks for the call set-up *progressive call control* is the most widely used. It means that each transit exchange selects the next section of the call route autonomously. In case of analogue exchanges, transit switches route the call, as it would have originated in that point. In digital exchanges also the incoming trunk on which the call arrived into the exchange may be taken into account for selecting the outgoing route.

In the practice progressive call control may be used efficiently. In principle, it has a drawback that when selecting the next section there is no information on the status of network sections beyond the next transit point. If no idle circuit is there, then the call will be lost. By proper network dimensioning, and/or by prioritising calls having seized several sections already, the risk of this is minimal.

Hierarchical routing systems

The traffic routing system is hierarchical, if for a given traffic route the set of overflow routes and the sequence in which those shall be used is given, irrespective of the destination of calls offered to it and of the routes already tested. This means that for the relation A-B the selection sequence of transit points T_1, T_2, ..T_i, …T_n is determined and for relation A-T_i only points after T_i may be used of the listed transit points (Figure 4.4.3). Practically, it may be easily realised when the frame architecture of the network is a hierarchical, multi-level tree, and in a transit node optional sections are to be selected in that sequence as their end point comes nearer on the route of the hierarchical frame.

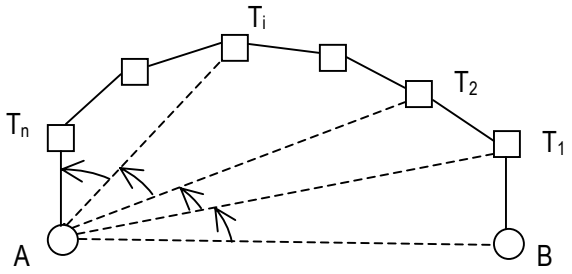


Figure 4.4.3: Hierarchical routing

A) Direct routing - if during the operation of the network the route to be selected is defined for each destination in the transit nodes, that shall be determined in the planning process selecting from the hierarchical section and the optional circuit groups. It is to be applied when the routing capability of exchanges is limited.

B) Alternative routing - if the exchanges are hunting for idle circuit in several outgoing routes in a given sequence and the first free route is selected. For this the appropriate capability of exchanges is needed. Since the alternative routing system seeks for further free route when tested optional routes are occupied, and the call is rejected only if also the last choice (hierarchical) circuit group is busy, it gives the possibility to apply so called *high usage circuit groups*. There is a high congestion on the high usage group, however the congested traffic overflows onto the next choice routes, at last to the hierarchical route. Only the hierarchical sections shall be dimensioned for low loss, and those carry the concentrated traffic of several relations,

together with the traffic of the hierarchical relation. Generally, this results in relatively high capacity and efficiently utilised circuit group.

Non-hierarchical routing systemes

If the network structure is meshed and there is no hierarchy defined among the exchanges, and routing changes depending on time periods or the state of the network, then dynamic and adaptive routing is used, respectively. These systems may be realised by digital (stored programme controlled) exchanges. Routing tables of the exchanges change by time periods or there is network controlling centre, which determines the routing tables based on monitoring and evaluating the state of the network.

References

[4.4.1] Wilkinson, R.I.. Theories for toll traffic engineering in the USA. Bell System Techn. Journ. Vol. 35. No.2. 1956.

[4.4.2] Larsson, T.: Technical long-range scheme of a country advaced in telecommunications. 3rd Telecom Forum, Part 2. No.311. Geneva, Sept. 1979.

[4.4.3] Lajtha, Gy., Borsos, K., Sallai, Gy., Papp, Z.: A change of the cost-effective network hierarchy. 3rd Telecom Forum, Part 2. No.354. Geneva, Sept. 1979.

[4.4.4] Sallai, Gyula (szerk.): Távközlő hálózatok forgalmi tervezése. KÖZDOK, 1980.

[4.4.5] Sallai, Gy., Papp, Z.: A statistical method for optimizing hierarchical networks. First Internat. Telecommunication Network Planning Symp. pp. 131-137. Paris, 1980.

[4.4.6] Sallai, Gy., Dely, Z.: Modular engineering and full grouping theorems for alternate routing networks. 10th Intern. Teletraffic Congress, No. 4.3 B-2. Montreal, 1983.

[4.4.7] Ash, G. R., Kafker, A.H., Krishnan, K.R.: Intercity dynamis routing architecture and feasibility. 10th Intern. Teletraffic Congress, No. 3.2.2. Montreal, 1983.

[4.4.8] Sallai, Gy., Dely, Z.: Dimensioning alternate routing networks with overload protection. 11th Intern. Teletraffic Congress, pp. 189-194. Kyoto, 1985.

[4.4.9] Roosma, H. A.: Optimization of digital network structure, 11th Intern. Teletraffic Congress,No. 2.4A-4. Kyoto, 1985.

[4.4.10] Girard, A., Cote, Y., Quimet, Y.: A comparative study of nonhierarchical alternate routing. 2nd Internat. Telecommunication Network Planning Symp. pp. 70-74, Brighton, March 1983.

[4.4.11] Sallai, Gy.: Computerized planning of telecommunication networks. Journ. Communications, Vol. XLII. March 1991. pp. 20-26.

[4.4.12] Dely, Z., Pauka, L., Balogh, I.: No.7-es jelzéshálózatok. OKTIG jegyzet, 1995.

[4.4.13] Dely, Z.: Jelzéshálózatok kialakítási lehetőségei. PKI Napok előadás, 1995.

[4.4.14] Kovács, Oszkár (szerk.): A keskenysávú ISDN kézikönyve. Távközlési Könyvkiadó, 1997.

4.5. Data communication networks

Sándor Mazgon, author

István Bartolits dr., reviewer

In data networks data terminal equipment change data with other data terminal equipment or network devices. The rules governing data interchange between data terminal equipment and between terminals and networking devices are called communication protocols. In that sense computer networks are data (communication) networks. This section deals with data communication networks and concentrating on the problematic of the data link layer.

Every data interchangein data communication networks are carried over data circuits by sending and receiving data sigals. Data terminal equipment have many relations to communicate over them, therefore we establish for a network of terminals some data connections to and from a set of intermediate nodes and between some pairs of nodes, instead of interconnecting all the terminals with each other directly. The network is interconnected, when each terminal can reach any other terminal via at least one route either directly or through some intermediate nodes. On each interconnecting link (direct circuit between two nodes) and on each access link (between the terminal and a node) there is at least one (possibly more multiplexed) data circuit. Those data circuits are circuit switched or packet traveling on those circuits are packet switched at the nodes to carry the end-to-end data flow between the communicating data terminals. At the circuit mode of operation a real connection is set up through the data circuits on the route before commencing data transfer, and is maintained for the full time of the call, and is released after the termination of the data transfer at the end of the call. Store and forward principle is governing the message/packet/cell switched and the frame relay networks, where data transfer is executed only on a single data circuit of the route at a time in sequence, step-by-step from the sending terminal to the destination. At each intermediate node there is a queue for messages/packets/cells waiting for the next data circuit to be transmetted over, if it is engaged with the transmission of other messages/packets/cells. The routing function and waiting in a store is executed for every message/packet/cell in

the node, this is very inefficient and resources consuming (processing power and storage). It means long delays if the system is based on full and some times very long messages, therefore (on the price of more processing power) messages are breaked into frames/packets of limited length or cells of constant length. This is the base of classifying data communication networks as packet switching, frame relay or cell relay data networks. Those technologies are relevant, may be adopted to the transmission equipment and to the switching equipment techniques as well.

4.5.1. Classification of data communications networks

There are many aspects of communication which can be used to classify data communicantions networks.

a. First is the industry or **community** for which the data network is the means of access to and carrier of telecommunication and value added services. There are networks and services freely available for the public, called **public** networks, others are restricted to some stated groups of users and called **private** networks. In most cases the usage of a **public** data network is paid by the user, some times at flat rate, in other cases for a fee depending on time, on bandwith and/or on distance of communication. Most of the **private** data networks provide data communication services without measuring and accounting traffic, others serve closed user groups on otherwise public networks, as if they were separated from the „public" users, this is called **virtual private data network**.

b. The second aspect is the **technology** used in the data network, which is somehow related to the geographical area covered. With the growth of a network the cost of bandwith, of delay or of recovery is growing. Therefore the technologies are continuously developed to fit better the needs, to get more robust and efficient solutions for the ever expanding scope of applications. Global data networks are provided to serve international needs and interconnectiong national and continental data networks using submarine cables and satellite links between gateway exchanges, while at the other end there are special data networks build into a single ship, an airplane, a vehicle, or an equipment interconnecting processors and processes of that very limited geographycal area. Most data networks are neither stand alone, nor specially separated from any other data network but are somehow interconnected to many other data networks, therefore the technologies used are depending also on the requirements of intercommunication.

c. The third aspect is the method of data communication. As data communications networks are designed and dimenzioned for the data traffic they are carrying, the method of data communication should also be taken into consideration. In that respect decisive factor is wether use is made of the connectionless mode of operation or of the connection-oriented mode of operation. For details see section 4.5.2.

Data communications networks are classified in the table 4.5.1 for another aspect: the network solutions aspect from the data communications users point of view. The traditional mode of classifying data communications networks are indicated on the table 4.5.2 for the range of transmission capability and in table 4.5.3 for the switching technique used.

Table 4.5.1. Data networking solutions

Comparison	Analog Dial-Up - PSTN	Switched 64	X.25	Point-to-Point Leased Lines	Frame Relay	xDSL	Cable	ISDN	ATM (Broadband ISDN)
Type	Circuit switched, public	Circuit switched, digital, private	Packet switched, public or private	Point-to-point, private	Packet switched, public or private	Circuit mode	Point-to-point, private	Packet/circuit switched, public	Cell switching technology
Bit rate	9.6-28.8 kb/s data (14.4 kb/s typical)	64 kb/s	<64 kb/s data	E1: 2 Mb/s, E3:34 Mb/s	64 kb/s - 2 Mb/s	64 kb/s - 140 Mb/s (varies by technology)	<140 Mb/s data and video	64-128 kb/s for BRI voice, video, and data 2 Mb/s for PRI voice, video, and data	2 Mb/s-622 Mb/s voice, video, and data (25-155 Mb/s typical)
Appli-cations	Voice and data on separate calls	Voice and data on separate lines	Protocol for host-to-host and for terminal-to-host	High-speed voice and data transmission for transaction-based environment applications and Internet access	Optimized for data transmission and point-to-point environments	Data dialtone (Internet access, remote LAN access), video dialtone (video conference, video on demand)	Consumer, on-line Internet and information access, LAN connections, video	Optimized for voice, data, and video integrated on a single digital line	Optimized for switching voice, data, video Switching , multiplexing technology
Number of Sites for Cost-Effec-tiveness	Unlimited	Unlimited	Unlimited	Few	Private leased line replacement; cost-effective for fewer fixed sites	Unlimited, but local access service only	Unlimited	Unlimited	Power users, early adopters; initially LAN only backbone applications
Strengths	Wide availability, any-to-any connectivity, low cost	Wide availability, any-to-any connectivity, moderate cost, uses standard telephone numbers, interoperates with ISDN	Wide availability, efficient for bursty traffic, any-to-any connectivity, automatic error detection and correction, security, standardized protocols	High speed; high degree of management, reliability, and security; standardized; direct connections to Internet	High speed, low latency, bandwidth-on-demand. easy scalability, standardized protocols, point-to-point connectivity	Simultaneous digital broadband and POTS services, supports multimedia service	High speed, existing infrastructure, fast call setup	High speed; digital data, voice, images, video on integrated line; fast call setup; secure, reliable, stable digital connectivity; efficient for bursty traffic; standardized protocols	Very high-speed; data, voice, images, video on integrated line; fast call setup; secure, reliable, stable digital connectivity; efficient for bursty traffic
Weak-nesses	Limited bandwidth, no efficiency gain supporting bursty versus continuous traffic, lacks multi-vendor management	Data only, limited bandwidth relative to ISDN, service providers de-emphasizing it, requires CSU/DSU equipment	Limited bandwidth since error detection limits speeds, increases cost; marginal for LAN interconnection	Fully meshed topologies very expensive	Expensive (compared to ISDN), requires a dedicated access line, not widely deployed, expensive and complicated to make moves and changes	Standards and infra-structure still under development, service area distance limitations, degree of data transport symmetry varies	Developmental stage only, voice on separate line, bandwidth split among users with no firewall capability, historically weak customer service and support, mostly one-way transmission	Not yet ubiquitous, tariff rates inconsistent, can be complicated to install and configure	Not yet widely available; standardization details still under development, current products expensive, proprietary products have multi-vendor compatibility problems due to lack of standards
Pricing Elements	Installation (varies); average monthly charge plus usage charge	Installation (varies); average monthly charge plus usage charges	Installation (varies); average monthly charge plus usage charges	Installation (varies); flat rate based on bandwidth and distance	Installation (varies); typically flat rate options, some carriers offer usage options	Installation (varies); average monthly charge plus usage charge	Installation (varies); average monthly plus usage charge (varies)		Customer-by-customer basis

Table 4.5.2. Classification of data networks by ranges

Public/Private	Public Data Network	Private Data Network				
Type		Wide Area Network	Metropolitan A. N.	Local Area Network	Small Area Network	Device Area Network
Abbreviation	PDN	WAN	MAN	LAN	SAN	DAN
Range	unlimited	≥ 100 km	≤ 100 km	≤ 10 – 35 km	≤ 100 – 500 m	≤ 10 – 35 m
Coverage	National and international PDNs are independent from the size of area covered and number of districts included	To interconnect many systems in distinct districts, mainly LANs and MANs	To interconnect many systems in a single large districts, mainly LANs	To interconnect many systems in a building or on a campus	Some devices are connected within an equipment or vehicle (ship, car, van, plane etc.)	Many units are inter-connected by a single medium within a system
Standard Interfaces	Interfaces specified in ITU-T Recommendations	World wide standard Interfaces	Standard interfaces like	Standard interfaces like ISO 8802 ie. IEEE 802	Proprietary interface of a group of vendors	Proprietary interface of a single vendor
Interopearbility	Any terminal shall be an Open System in OSI sense	Many Open Systems communicate in OSI sense	Many Open Systems com-municate e.g. in OSI sense	Many Open Systems com-municate on common base	Groups of devices interact by proprietary rules and processes	Devices interact by proprietary rules and processes of the vendor
Standard Protocols	Protocols specified by ITU-T Recommendations See Fig 1./X.220	Protocols specified by international standars and agreements	Mainly internationally standardised protocols	Primarly publicly available standard protocols	Proprietary protocols of a group of vendors	Proprietary protocols of a single vendor

Table 4.5.3. Comparison of switching technologies used for data communication:

Characteristics	Circuit switching	Message switching	Packet switching 1 virtual call	Packet switching 2 datagram	frame relay and switching	cell switching
Line condition of the end sections between communications	Off-hook	idle condition (always ready to transmit)	idle condition (always ready to transmit)	idle condition (always ready to transmit)	idle condition (always ready to transmit)	idle condition (always ready to transmit)
Call set up	The whole end-to-end real path is established	No call set up	End-to-end virtual path is allocated	No call set up	Dynamic call set up	Dynamic call set up
Resource allocation before and during call set up	call detection, routing, switching,	buffers are virtually allocated for each message to be transfered	buffers are virtually allocated for each packet to be transfered	buffers are virtually allocated for each datagram to be transfered	buffers are virtually allocated for each frame to be transfered	buffers are virtually allocated for each cell to be transfered
Transfer channel occupation	end-to-end during the call	section by section a message at time	section by section a packet at time	section by section a datagram at a time	section by section a frame at a time	section by section a cell at a time
Occupied channel bandwith	constant bandwith in all section	different bandwith (dinamically allocated on trunk lines)	different bandwith (dinamically allocated on trunk lines)	different bandwith (dinamically allocated on trunk lines)	different bandwith (dinamically allocated on trunk lines)	different bandwith (dinamically allocated on trunk lines)
Message store at any node	no store for messages	complete message is stored before processing (for routing)	store and forward operation is performed on each packets in sequence	store and forward operation is performed independently on each packets	store and forward operation is performed on each frames in sequence	store and forward operation is performed on each cells in sequence
Length of data unit	arbitrary	limited by the store	maximal packet length is network dependent	maximal packet length is network dependent	maximal frame length is network dependent	length of a cell is standardized (48+5 octets)
Delay variation	no variation of transfer delay	waiting and processing delays have strong variations	waiting and processing delays have strong variations	waiting and processing delays have strong variations	waiting and processing delays have strong variations	waiting and processing delays have strong variations
Traffic pricing elements	call duration and distance	data volume transferred or message length and urgency	volume of data in segmenst, and distance	volume of data in segmenst, and distance	volume of data in segmenst, and distance	number of cells, and distance

4.5.2. Operation of of data communication networks

Data communication networks are service oriented. Private data network provide communications services for closed user groups of their own. Public data networks (and ISDNs) provide data transmission services in one or more of the form (see ITU-T Recommendations X.7 and F.600):

circuit switched data transmission service (CSDTS)
packet switched data transmission service (PSDTS)
frame relay data transmission service (FRDTS)
leased circuit data transmission service (LCDTS).

Operation of data communications network depend on their status (private or public) on the services provided and other aspects dscribed below.

Connectionless vs. connection-oriented operation

In connectionless data communication there is no sequence number and acknowledgment scheme. No sequence numbers means that any packet is in the network a single datagram per se and is independent from any other datagrams. This kind of data communications does not require that a connection be established before data can be sent. Analogous to submit a telegram or mailing a letter (you may have addressed it incorrectly, but you don't find out about the problem until after the network can not deliver it). Usually the higher-layer protocols provide more robust error detection and correction than for connection-oriented communication. UDP protocol is connectionless.

In connection-oriented data communication there are sequence numbers and an expected acknowledgment process. If acknowledgments are not received, a retransmission occur. This way the connection-oriented protocol can guarantee the sequenced delivery of data. If a protocol use sequence numbers then it can be viewed as a ‘connection-oriented’ and as a ‘reliable’ protocol. This kind of data communications requires that a (real or virtual) connection first be established (*caller* calls i.e „requests a connection”, *callee* „accepts the connection”) before data can be exchanged. Analogous to a telephone call (you can't just pick up the phone and start talking before you get the dial tone, you dialled and waited until the callee ringed, was available, was taking off the phone and is "listening"). The connection-oriented

mode of communication usually is more reliable than a connectionless one. The X.25 networks and the TCP protocol are connection-oriented.

Protocols for data communication networks

Until recently there was a distinct difference between the sorts of systems developed for data transmission over general-purpose, public telecommunications networks and those developed specifically for computer-to-computer communications as network of (private) networks. Increasingly, however, the distinction between the Open System Interconnection (**OSI**) standards developed by member bodies of the International Standards Organization (**ISO**)[3] and the International Telecommunication Union (**ITU**)[4] and computer-to-computer communications protocols developed for the Internet by the Internet Engineering Task Force (**IETF**)[5] have become blurred. As a result the two systems can, for most data interchange purposes, now be seen as component parts of the global information infrastructure.

However, there is no direct symmetry between the OSI application standards and the Internet application level protocols. Differences in their technical nature are closely related to the very different fundamental principles which underlie the OSI Reference Model and the Internet protocol suite, and the very different manners in which OSI standards and Internet protocols have been developed:

- OSI application standards are based on an upper layer architecture and discrete modules. The three upper layer stacks provide *full interworking flexibility* (at a cost of complexity). The OSI upper layers are in principle independent of the lower OSI layers as well as each protocol in every layer can be replaced by an other protocol without disturbing the communication of the other layers.
- Internet application protocols are written from the ground up. TCP/IP based application services such as FTP and TELNET are plugged directly into the Transport services.

It is in principle possible to run OSI application services over any lower layer stack, including over TCP/IP. At a more general level, any consideration about the relationship between OSI and TCP/IP needs to cover two distinct, albeit inter-related, aspects -- the protocols for the upper layers on the one hand, and the transport

[3] http://www.iso.ch

[4] http://www.itu.int; http://www.itu.ch

[5] http://www.ietf.org; http://www.ietf.cnri.reston.va.us

protocols on the other. The co-existence and interoperability of OSI, TCP/IP and proprietary protocol suites has been made possible and in recent years resulted in a number of specifications on multi-protocol operation, which aim to provide an "any-to-any" solution (any application service over any lower layer protocol stack).[6]

4.5.3. Conventions of data networks

For data communication networks there are some fundamental conventions as conditions of interoperability as first targets of standardization. One of the series of such standards are the mechanical interface standards used worldwide. The same way the operation on a network must be based on the unique global interpretation of any bit and of any data signal or control signal crossing the border of the network. This convention is given in the table 4.5.4 below:

Order of bits in data transmission

Another convention shall be the order of bits on the line. Seen from the frame level of communication: the addresses, commands, responses and sequence numbers are transmitted (to the line) with the low-order bit first (for example, the first bit of the sequence number that is transmitted have the weight 2^0). The FCS, the frame checking sequence is transmitted to the line commencing with the coefficient of the highest term (bit position 16).(See ITU-T Recommendation X.25 and X.75)

A facsimile page is composed from a set of page-wide stripes of image data, which are coded independently. The stripes are transmitted sequentially from the top to the bottom of the page. The bit order of the coded data at the output of a FAX encoder is described as most significant bit first, i.e. bits are packed into octets starting at the most significant bit. All multi-octet values shall be interpreted in a most-significant-first manner: the first octet of each value is most significant, and the last octet is the least significant. In a data communication system the coded data are transmitted least significant bit first for each octet on the communication line in a bit stream of least to most significant bit order. The order of bit transmission, octet by octet, from an originating FPAD to the physical line of the packet network is the same

[6] A summary of these available specifications can be found in the EWOS Technical Guide ETG 053, *Overview of approaches to multiprotocol coexistence and convergence in support of the Transport Service*. [http://www.ewos.be]

	SIGNIFICANT CONDITION		NOTE
CCITT notation	**condition A**	**condition Z**	Basic convention
Symbol	**0**	**1**	See CCITT Recomm. V.1, V.4, V.10, V.11 and X.4
Binary digit	**0**	**1**	
Bit value	**0**	**1**	
Start-stop code	**start-signal**	**stop-signal**	See CCITT Recomm. V.1, U.1--U.12, X.20, X.21, X.70 and X.71
Start-stop code	**space-element**	**mark-element**	
Telex switching	**line available**	**line idle**	
Amplitude modulation	**tone-off**	**tone-on**	OOK (On-Off-Keying), R.31
Frequency modulation	**high frequency**	**low frequency**	Frequency keying, R.35
Phase modulation with reference phase	**opposite phase to the reference phase**	**reference phase**	
Differential two-phase modulation	**no phase inversion**	**inversion of phase**	Relative to the previous element
Paper tape	**no perforation**	**perforation**	
DATA Circuit	**0**	**1**	See interface circuits in the R-, S-, T-, V- and X- series of the CCITT Recommendations
CONTROL Circuit	**ON**	**OFF**	
TIMING Circuit	**ON**	**OFF**	
POTENCIAL	**positive**	**negative**	on interface circuits
Receiver significant levels	**≥ + 0,3 V**	**≤ - 0,3 V**	V.10 and V.11 (A relative to B)
Receiver significant levels	**≥+ 3 V**	**≤ - 3 V**	V.28 (A relative to common return)
NAMING CONVENTION IN THE TWO-CONDITION CODE			
English	**space**	**mark**	
French	**travail**	**repos**	
Spanish	**trabajo**	**reposo**	
Hungarian	**jel**	**szünet**	
German	**Zeichen**	**Trenn**	(Schritt = element)
German	**Arbeits**	**Ruhe**	(Zustand = condition)

Table 4.5.4. EQUIVALENCE BETWEEN BINARY NOTATION SYMBOLS AND THE SIGNIFICANT CONDITIONS OF A TWO-CONDITION CODE (as of CCITT Recommendation V.1)

According to international agreements and practice the naming conventions are not simple translations. The continental European and the English naming conventions differ from translations.

as the bit order from the local G3 facsimile equipment to the originating FPAD on the PSTN line. The receiving FAX will reorder the bits received from the line and treat the coded data as most significant bit first, so if a decoder is reading a sequence of bits from the octet-stream, it reads first the most significant bit of the first octet, the next most significant bit, and so on, then proceed to the next octet [See ITU-T Recommendation T.4 (G.3 fax), T.85, and X.39 (FPAD)].

4.5.4. Elements of Data Network

X.25 user-network interface (UNI) for the packet switched public data networks

X.25 is the recommended interface between the connection oriented, packet switched, public data communication network and the packet mode data terminal equipment and includes the lowest three layer protocols of the 7 layer OSI model. It was standardized by the CCITT of the ITU (forerunner of the ITU-T, which is now maintaining and developing all the X-series of recommendations) for data communication services offered by common carriers. It defines how data connections by protocol exchanges between end user devices (DTE, „data terminal equipment") and network devices (DCE, „data circuit terminating equipment") are established, maintained and released. X.25 was designed to operate effectively regardless of the type of systems connected to the network. The development of the X.25 recommendation was initiated by the common carriers in the early 1970s in paralel to the first private and research networks which later became part of the Internet.

In X.25 the physical interface may be according to ITU-T Recommendation X.21 or X.21 bis for dedicated lines, X.31 for (switched) ISDN access and X.32 for (switched) PSTN, ISDN and CSPDN access. Transmission may be bit-synchorous or sharacter-by-character for start-stop characters. The data link layer's protocol is a Link Access Procedure Balanced (**LAPB**) of the High-level Data Link Control Procedures (**HDLC**) class and may be of a Single Link Procedure (**SLP**) or a MultiLink Procedure (**MLP**). In an MLP the SLPs are used independently on each physical circuit (if many). The network layer protocol is a Packet Layer Protocol (**PLP**) supporting virtual circuit services, i.e. both permanent virtual circuit (**PVC**) and virtual call (**VC**, virtual call, which is the same as a **SVC**, switched virtual circuit) services. On a single data link multiple virtual circuits may be established, which are identified by their logical channel group number and a logical channel number. The logical channels are means of duplex transmission.

X.75 network-network interface (NNI) for the packet switched public data networks

Those networks, in which the user network interface conform the ITU-T Recommendation X.25, use the network-network interface specified in the ITU-T

Recommendation X.75. It consists of the symmetrized variants of the naturally asymmetric procedures of X.25 and is extended by the means to interchange control information between networks.

the PAD/FPAD

The packet assembler/disassembler (PAD) is a device common in X.25 networks. PADs are used when a DTE (data terminal equipment), such as a character-mode terminal, is too simple to implement the full X.25 functionality. The **PAD** is located between a **DTE** and a **DCE** (data circuit terminating equipment), and performs the functions: buffering, packet assembly and disassembly, set up and release of virtual calls, character handling, deletion and insertion of start and stop bits from and to the start-stop **DTE**. The **PAD** buffers data sent to or from the **DTE** device, assembles outgoing data into packets, adds to each an X.25 header and forwards them to the **DCE**, as well as disassembles incoming packets, removes the X.25 header and forwards the received data to the **DTE**. Between Group-3 (**G3**) facsimile terminals special **FPAD** devices provide the **PAD** service. (See ITU-T Recomendations X.3 and X.5 respectively.)

LAPB, the protocol of data link layer is of HDLC type

Data link control in X.25 networks includes procedures for error recovery, special rules to use FRMR (framereject), REJ (reject) and SREJ (select reject) frame, procedures for collision detection and resolution, procedures for flow control by window mechanism, procedures for resolving exception conditions, and procedures to handle system parameters (e.g. k, the maximum number of sequentially numbered I frames outstanding, i.e. unacknowledged). May also include a Multi Link Procedures (MLP) to accept packets from the Packet Layer and to distribute those packets across the available Single Link Procedures (SLPs), and resequencing the packets received from the SLPs for delivery to the Packet Layer, respectively, and to resolve error and exceptional conditions, to execute multilink window mechanism.

Frame type, frame structure and frame address in LAPB

The LAPB used in X.25 assumes a direct physical link betwen a single DTE and a single DCE, where frames are interchanged between the two. Each frame is either a command or a response:

Information type: Information (I) frame is allways a command.

Supervisory types: Receive Ready (RR), Receive Not Ready (RNR) and Reject (REJ) frames may be commands or responses, Selective reject (SREJ) frame may only be a response.

Unnumbered types: Set Asynchronous Balanced Mode (SABM), Set Asynchronous Balanced Mode Extended (SABME), Set Mode (SM) and Disconnect (DISC) frames may only be commands, Unnumbered Acknowledgement (UA), Disconnected Mode (DM) and Frame reject (FRMR) frames may only be responses.

All transmissions are in frames. Each frame consists of an opening **flag**, an **address** field (one octet), a **control** field, optional **information** field, a **frame checking sequence** (FCS) and a closing **flag**. A single **flag** (the 01111110 sequence) may be the closing flag for one frame and the opening flag for the next frame. The **address** field identifies the intended receiver of a command frame and the transmitter of a response frame. The three types of **control** field formats are: numbered information transfer (**I**) format, numbered supervisory functions (**S**) format and unnumbered control functions (**U**) format. The **I format**, used for information transfer, contains sequence numbers, which may or may not acknowledge **I** frames, and an independent **P** bit. The **S format**, used for acknowledging and for requesting retransmission or a temporary suspension of transmission, contains as sequence number the expected send sequence number of the next received I frame, two supervisory **S** bits and an independent **P/F** bit. In command frames, the **P/F** bit is referred to as the **P** bit (1=poll, i.e. soliciting a response) and in response frames, it is referred to as the **F** bit (1=final, i.e. a solicited response). The **sequence numbering** may have three modes of operation: the basic (modulo 8 operation), the extended (modulo 128 operation) and the super (modulo 32 768 operation), and influences the length of the control field containing the sequence numbers: one octet (basic), two octets (extended) and four octets, respectively. The **U format,** used to provide additional data link control functions, contains no sequence numbers, has therefore control field length of one octet, does include 5 modifier function bit and a P/F bit.

The **FCS** field is a specially calculated 16-bit sequence checking all the bits between two flags excluding the inserted extra bits for synchronization and the stuffing zeros.

The address used in the frame (for the second layer addressing) depends on the frame type, is it a command or a response, and on the link procedure: is it an SLP or an MLP:

	Command	Response
	(your address)	(my address)
from DCE to DTE	SLP: address **A** MLP: address **C**	SLP: address **B** MLP: address **D**
from DTE to DCE	SLP: address **B** MLP: address **D**	SLP: address **A** MLP: address **C**

Table 4.5.5

HDLC family of procedures (protocols for the 2nd layer in OSI):

HDLC high level data link control procedures (for ISO-OSI, see ITU LAPs)

LAPB link access procedure balanced (for datacom., see ITU-T X.25/X.75)

LAPD link access procedure on the D-channel (see ITU-T Q.921)

LAPF link access procedure for frame mode bearer services (see Q.922)

LAPM link access procedures for modems (see ITU-T V.series of Recs.)

LAPX half duplex link access procedures (see ITU-T T.71 + X.32)

PLP the protocol of the third layer

The **PLP** operates in five distinct modes: *call setup*, *data transfer*, *idle*, *call clearing*, and *restarting*. *Call setup* and *call-clearing mode* is used on a per virtual circuit basis and is used only with **SVCs** (switched virtual circuits) to establish **SVCs** between **DTE** devices, and to terminate **SVCs** between **DTE** devices, respectively. *Data-transfer mode* is used with both **PVCs** (permanent virtual circuits) and **SVCs** for transferring data between two **DTE** devices across a virtual circuit. In this mode, **PLP** handles segmentation and reassembly, bit padding, error and flow control. *Idle mode* is used when a virtual circuit is established but data transfer is not occurring. *Restarting mode* is used to synchronize transmission between the DTE and the DCE, i.e to initialize or reinitialize the packet layer DTE/DCE interface. The **restart procedure** simultaneously clears all the virtual calls and resets all the permanent virtual circuits at the DTE/DCE interface. The **interrupt procedure** allows a DTE to

transmit data to the remote DTE, without following the flow control procedure applied to the normal *data* packets. Only *data* packets contain a sequence number. The sequence numbering scheme of the packets is performed modulo 8 (basic), modulo 128 (extended) or modulo 32 768 (super extended) and is the same for both directions of transmission and is common for all logical channels. A **window mechanism** is provided to perform flow control. The standard window size W is 2 for each direction of data transmission, other window size may be selected for a period of time for each permanent virtual circuit and may be negotiated on a per call basis for each virtual call with the *flow control parameter negotiation* facility. It is a network option to provide **delivery confirmation**. A **reset procedure** is used to reinitialize the virtual call or permanent virtual circuit. Other procedures are provided to resolve and to recover from fault conditions, collisions, exceptional situations etc.

PLP packets and PLP addresses

Every packet transferred across the DTE/DCE interface consists of at least three octets: a general format identifier (**GFI**), a logical channel identifier (**LCI**) and a packet type identifier (**PTI**). User data in octetts are included in data packets. For modulo 32 768 operation, a protocol identifier octet is contained in the first octet of each packet. **GFI** identifies whether the packet carries user data or control information, what kind of windowing is being used, whether delivery confirmation is required, and wether **TOA/NPI address format** is used. **LCI** identifies the virtual circuit by its logical channel number and logical channel group number across the local DTE/DCE interface. **PTI** identifies the packet as one of 39 different packet types. **User Data** contains encapsulated upper-layer information. Other packet fields are appended as required.

The call set-up and clearing packets contain an address block. This address block has two possible formats. The first format, the **non-TOA/NPI address format**, can accommodate addresses conforming to ITU-T Recommendations X.121 and X.301 whose length (including possible prefixes and/or escape codes) is not greater than 15 digits. The second format, is the **TOA/NPI address format**, can accommodate addresses conforming to Recommendations X.121 and X.301 whose length is greater than 15 digits and can also be used to carry an alternative address in the called DTE address field. The address block of the TOA/NPI format makes

possible, that each DTE address field, when present, has three subfields: a subfield to specify the **type of address** (**TOA**), a subfield to specify the **Numbering Plan Identification (NPI)** and the address digits subfield.

4.5.5. Local Area and Metropolitan Area Networks (LAN/MAN)

Local area network originally is a network build on a single, common medium, which is accessed by all of the communicating terminal stations. With special devices this medium is expanded and extended to buid a complex system of interconnected media, where different access methods, different topologies, different management systems can communicate with each other on the base of a common policy. The system is extendable to a larger area too, which is then the MAN, the metropolitan area network. The most widely known LAN technology is obviousely the Ethernet.

"Ethernet" is the term that is applied to a number of very different data link implementations. To refer to the original "Ethernet", invented by Robert Metcalf, you mean the implementation of DEC, Intel, and Xerox of Version 1 or Version 2 Ethernet before the IEEE established the 802.3 standards in 1984. The IEEE 802 group of standards is the base of allmost any LAN architecture and technology and bult up the way how the OSI layering principles are practicized in the two lower layers for LANs. The American national standards of the IEEE 802.x group became global standards as ISO/IEC 8802.x by JTC 1 of ISO and IEC. They include beyond Ethernet also other technologies, like token ring, token bus, etc. see table 4.5.6. Description

802/	Overview & Architecture	IEEE 802-1990	ISO/IEC 8802
1	**LAN/MAN Bridging & Management**	IEEE 802.1-1993 IEEE 802.1G, 1998	ISO/IEC 8802.1 ISO/IEC TR11802-x ISO/IEC 15802-x ISO/IEC15802-5:1998
2	**Logical Link Control**	IEEE 802.2-1998	ISO/IEC 8802-2:1998
3	**CSMA/CD Access Method**	IEEE 802.3-2000	ISO/IEC 8802-3:2000
4	**Token-Passing Bus Access Method**	IEEE 802.4-1990 (R1995)	ISO/IEC 8802-4:1990
5	**Token Ring Access Method**	IEEE 802.5-1998	ISO/IEC 8802-5:1998
6	**DQDB Access Method**	IEEE 802.6-1994	ISO/IEC 8802-6:1994
7	**Broadband LAN**	IEEE 802.7-1989 (R1997)	
10	**LAN/MAN Security**	IEEE 802.10-1998	
11	**Wireless**	IEEE 802.11-1999 IEEE 802.11a-1999	ISO/IEC 8802-11: 1999 ISO/IEC 8802-11/Amd 1:2000(E)
12	**Demand Priority Access Method**	IEEE 802.12-1998	ISO/IEC 8802-12:1998
14	**Wireless Personal Area Networks**	IEEE 802.14	
15	**Broadband Wireless Access**	IEEE 802.15	
16	**Standard Air Interface**	IEEE 802.16	
17	**Resilient Packet Ring**	IEEE 802.17	

Table 4.5.6

include the non 802 standard technologies (FDDI & CDDI), which are important for LAN/MAN implementations.

LAN/MAN **media access** protocols typically are: *carrier sense multiple access collision detect* (CSMA/CD) and *token passing*. Token-passing media-access scheme are Token Bus/IEEE 802.4, Token Ring/IEEE 802.5 and FDDI.

FDDI and CDDI

The **Fiber Distributed Data Interface (FDDI)** is a 100-Mb/s token-passing, dual-ring LAN using fiber-optic cable. FDDI is frequently used as high-speed backbone technology because of its support for high bandwidth and greater distances than copper. A new specification, called **Copper Distributed Data Interface (CDDI**) has emerged to provide 100-Mb/s service over copper. CDDI is practically the implementation of the FDDI protocol over twisted-pair copper wire.

FDDI uses a dual-ring architecture with traffic on each ring flowing in opposite directions (called *counter-rotating*). The dual-rings consist of a primary and a secondary ring. During normal operation, the primary ring is used for data transmission, and the secondary ring remains idle. The primary purpose of the dual rings, is to provide superior reliability and robustness. Finally, FDDI allows two kilometers between stations using multi-mode fiber, and even longer distances using a single mode meaning some 100 kilometres for an FDDI network of 1000 stations. CDDI supports distances of about 100 meters from desktop to concentrator.

FDDI is a collection of four separate specifications. FDDI's four specifications are the **Media Access Control (MAC**), **Physical Layer Protocol (PHY**), **Physical-Medium Dependent (PMD**), and **Station Management (SMT**). The **MAC** specification defines how the medium is accessed, including frame format, token handling, addressing, algorithms for calculating cyclic redundancy check (**CRC**) value, and error-recovery mechanisms. The **PHY** specification defines data encoding/decoding procedures, clocking requirements, and framing, among other functions. The **PMD** specification defines the characteristics of the transmission medium, including fiber-optic links, power levels, bit-error rates, optical components, and connectors. The **SMT** specification defines FDDI station configuration, ring

configuration, and ring control features, including station insertion and removal, initialization, fault isolation and recovery, scheduling, and statistics collection.

Unique characteristics of FDDI is that FDDI defines three types of devices by which to connect FDDI devices: single-attachment station (**SAS**), dual-attachment station (**DAS**), and a **conc**entrator. An **SAS** attaches to only one ring (the primary) through a concentrator. One of the primary advantages of connecting devices with SAS attachments is that the devices will not have any effect on the FDDI ring if they are disconnected or powered off. Each **DAS** has two ports, designated A and B. These ports connect the DAS to the dual FDDI ring. Therefore, each port provides a connection for both the primary and the secondary ring. Devices using DAS connections will affect the ring if they are disconnected or powered off. An FDDI concentrator, the so called a *dual-attachment concentrator* [**DAC**]) is the building block of an FDDI network. It attaches directly to both the primary and secondary rings and ensures that the failure or power-down of any **SAS** does not bring down the ring. This is particularly useful when PCs, or similar devices that are frequently powered on and off, connect to the ring. FDDI's primary fault-tolerant feature is the dual ring. If a station on the dual ring fails or is powered down, or if the cable is damaged, the dual ring is automatically *wrapped* (doubled back onto itself) into a single ring. When the ring is wrapped, the dual-ring topology becomes a single-ring topology. Data continues to be transmitted on the FDDI ring without performance impact during the wrap condition. When a single station fails, devices on either side of the failed (or powered down) station wrap, forming a single ring. Network operation continues for the remaining stations on the ring. When a cable failure occurs, devices on either side of the cable fault wrap. Network operation continues for all stations, but FDDI truly provides fault-tolerance against a single failure only. When two or more failures occur, the FDDI ring segments into two or more independent rings that are unable to communicate with each other.

Addressing

Datacasting technique used in data communication networks fall into three classes: *unicast*, *multicast* and *broadcast*. In each class of transmission a single packet is sent to one or more nodes. In an *unicast transmission*, a single packet is sent from the source to a single destination on a network. First, the source node

addresses the packet by using the unique address of the destination node. The packet is then sent into the network, and finally, the network passes the packet to its destination. A *multicast transmission* consists of a single data packet that is copied and sent to a specific subset of nodes on the network. First, the source node addresses the packet by using a multicast address. The packet is then sent into the network, which makes copies of the packet and sends a copy to each node that is part of the multicast address. A *broadcast transmission* consists of a single data packet that is copied and sent to all nodes on the network (*general broadcast*) or to a given number of groups (*selective broadcast*). In these types of transmissions, the source node addresses the packet by using the broadcast address. The packet is then sent into the network, which makes copies of the packet and sends a copy to every node on the network. Simple LAN networks use these addressing on their single medium such, that the selection is made at the destination node in the receiver process as all the nodes receive all the packets as they appear on the medium.

Topology

LAN **topologies** define the manner in which network devices are organized. Four common LAN topologies exist: bus, ring, star, and tree. These topologies are logical architectures, but the actual devices need not be physically organized in these configurations. Logical bus and ring topologies, for example, are commonly organized physically as a star. A bus topology is a linear LAN architecture in which transmissions from network stations propagate the length of the medium and are received by all other stations. Ethernet/IEEE 802.3 networks, including 100BaseT, implement a bus topology. A ring topology is a LAN architecture that consists of a series of devices connected to one another by unidirectional transmission links to form a single closed loop. Both Token Ring/IEEE 802.5 and FDDI networks implement a ring topology, or by two-directional links forming two independent or interdependent loops.. A star topology is a LAN architecture in which the endpoints on a network are connected to a common central hub, or switch, by dedicated links. Logical bus and logical ring topologies are often implemented physically in a star topology. A tree topology is a LAN architecture that is identical to the bus topology, except that branches with multiple nodes are possible in this case without allowing any looping..

LAN/MAN special devices

A *repeater* is a physical layer device used to interconnect the media segments of an extended local network. A repeater essentially enables a series of cable segments to be treated as a single cable. Repeaters receive signals from one network segment and amplify, retime, and retransmit those signals to another network segment. These actions prevent signal deterioration caused by long cable lengths and large numbers of connected devices. Repeaters are incapable of performing complex filtering and other traffic processing. In addition, all electrical signals, including electrical disturbances, noises and other errors, are repeated and amplified. The total number of repeaters and network segments that can be connected is limited due to timing and other issues.

A *hub* is a physical-layer device that connects multiple user stations, each via a dedicated cable. Electrical interconnections are established inside the hub. Hubs are used to create a physical star network while maintaining the logical bus or ring configuration of the LAN. In some respects, a hub functions as a multiport repeater.

References

[4.5.1] Béla Sebestyén: Local area networks (In Hungarian. Helyi számítógép-hálózatok), Publisher: Műszaki Könyvkiadó. Budapest.1987

[4.5.2] Sándor Balogh, Jenő Berkes, László Kovács: Telematic services of computer telecommunication. (In Hungarian. A számítógépes távközlés telemaikai szolgálatai.) Publisher: OMIKK LSI Alkalmazástechnikai Tanácsadó Szolgálat, Budapest. 1988

[4.5.3] Jenő Berkes, László Gonda, Károly Szabó, Attila Verebélyi: Data transmission for computer users. (In Hungarian. Adatátvitel számítógép felhasználóknak.) Publisher: Ipari Informatikai Központ (Informatik), Budapest. 1989

[4.5.4] Uyless D. Black: Data Networks, concepts, theory and practice. Prentice-Hall. 1989

[4.5.5] Tannenbaum, Andrew S.: Computer Networks. Panem-Prentice-Hall. New Jersey 1996. (Third edition in Hungarian: Számítógép-hálózatok. Harmadik kiadás magyar változata. Budapest.1999. Edited by: dr. József Harangozó.)

[4.5.6] „Free on-line dictionary of computing" http://www.nightflight.com/foldoc/contents/networking.html,

see inter alia *connection oriented* (e.g. TCP) or *connectionless* (e. g. UDP).

[4.5.7] For LAN and TCP/IP (Internet technology) readings see:

Lecture Information by Godred Fairhurst: EG3561/EG3567 Communications Engineering.
http://www.erg.abdn.ac.uk/users/gorry/eg3561/road-map.html,
http://www.erg.abdn.ac.uk/users/gorry/eg3561/index.html

[4.5.8] Technical tutorials, guidlines and "webtorials" to be found:

General: http://www.internet.com/sections/downloads.html
http://webreference.com/internet/tutorials.html
webopedia: http://www.webopedia.com/
http://www.pcwebopedia.com/
nanotech planet: http://www.nanotechplanet.com/
Internet lists: http://www.internet.com/sections/lists.html
Internet technology: http://www.internet.com/sections/it.html
Internetdotcom: http://www.webreference.com
datacom: http://www.alliancedatacom.com/
Internet-standards: http://webreference.com/internet/standards.html
http://www.wssn.net/WSSN/
info-superhighway: http://www.netsquirrel.com/index.html

Abbreviations

ITU-T, CCITT, ISO, IEEE, IEC, IETF, EWOS (European Workshop for Open Systems), ETO, OSI, PSTN, E1, E3, kb/s, Mb/s, ISDN, xDSL, FR, ATM, B-ISDN, PDN (public data network), WAN, MAN, LAN, SAN (small area network), DAN (device area network), CSDTS, PSDTS, FRDTS, LCDTS, TCP, IP, UDP, FTP, TELNET, FCS, PAD, FPAD, AM, OOK, FM, UNI, NNI, DTE, DCE, CSPDN, PSPDN, HDLC, SLP, MLP, PLP, VC, PVC, SVC, FRMR, REJ, SREJ, I, RR, RNR, SABM, SABME, SM, UA, DISC, DM, LAP, LAPB, LAPD, LAPF, LAPM, LAPX, FCS, S, U, P, F, GFI, LCI, PTI, TOA, NPI, FDDI, CDDI, CSMA/CD, DQDB, MAC, PHY, PMD, SMT, CRC, SAS, DAS, DAC, PC.

Appendix to section 4.5: Terms and definitions

Data network

Telecommunications network planned and optimized to provide data transmission services, the main capability of which is to transfer data efficiently among data stations (of data terminal equipment), or among data terminal equipment and other telecommunications networks, parts of such networks or other equipment, each of which is connected to the data communication network by network interconnection device.

Related terms: data communication network, data transmission network

Open Network Provision (ONP - EU)

"Open network provision conditions" means the conditions, harmonized according to the provisions of the **ONP** Framework Directive, which concern the **open and efficient access to** public telecommunications networks and, where applicable, public telecommunications services **and the efficient use of** those networks and services (*see EU-directives*).

Information

Any kind of knowledge, that is exchangeable among users, about things, facts, concepts in a universe of discourse. Although information will necessarily have representation form to make it communicable, it is the interpretation of this representation (**the meaning**) that is relevant in the first place. [ISO/IEC 10746-2:1996 | ITU-T X.902 (1996) §.3.2.5]. Information is the meaning that a human assigns to data by means of known conventions used in their representation.

Data

The representation form of information dealt with by information systems and users thereof. [ISO/IEC 10746-2:1996 | ITU-T X.902 (1996) §.3.2.6] Representation of facts, concepts or instructions in a formalized manner suitable for communication, interpretation, or processing by humans or by automatic means. Any representation such as characters or analogue quantities to which meaning is or might be assigned.

Data transmission

That part of data communication process which provides for the transfer of data between locations mostly by electromagnetic means, where data are carried by data signals.

Data communication network

Telecommunications network planned and optimized to provide data transmission services, the main capability of which is to transfer data efficiently among data stations (of data terminal equipment), or among data terminal equipment and other telecommunications networks, parts of such networks or other equipment,

each of which is connected to the data network by network interconnection device (IWU, interworking unit).

Related terms:: data network, data transmission network, computer netwok, computer communication network.

4.6. Special purpose networks

András Bély, István Jutasi, authors

István Bartolits, reviewer

Generally, telecommunication is a kind of service providing two-way information-exchange without turning out a product. It is also a personal service inasmuch as everybody can own a device meeting his demands and providing access to the required service.

However, there is also a less well-known special field of telecommunication that is usually impersonal but offers, as an integral part of a given technological field or system, much more benefit. This latter is the field of technological telecommunication, and the network itself is called special purpose network (also called technological network). The most important feature of special purpose networks is the availability, further a series of specific services that may be offered as required by the technology in question.

4.6.1. Vehicular services

On land, economical and safe transportation has been made possible by modern railway traffic. The velocity increase of the railway traffic has been rendered possible – though indirectly – by the Morse telegraph line between Washington and Baltimore, put into operation in the year of 1844. From this time onwards, telecommunication became an inseparable part of railway traffic.

In maritime navigation, a decisive change was introduced by Marconi's invention, the wireless telegraph. Later, this was followed by radio communication, and recently, satellite communication.

Railway communiation

In railway networks, communication is applied for promoting the safety of traffic control and for providing infnormation for the railway traffic and maintenance service. The information sent over the telegraph line, the ”telegram” serves also as a

written document, while the telephone, on the other hand, is used to provide verbal instructions and reports for the traffic maintenance personnel.

In the beginning, railway networks were realized by telegraph lines, in parallel with the railway track, utilizing bronze wires tightened on a row of wooden columns, later utilized also for telephone transmission. The railway track could later be utilized by the post office for building its own network. Consequently, along the track, one side was utilized for the railway technological network while along the other side, the public postal telephone network was operated.

Railways are usually operated by powerful organizations comprising operating and maintenance departments, the former handling the transport activities and the direct traffic arrangements. The communication devices directly related to the operating technologies involved with given stations, nodes and railway tracks are under the disposal of the operational service. These devices comprise the special purpose and private telephone links, dispatcher systems, radio districts and systems, public announcement systems for giving instructions, the electric clock system, voice recording apparatus, acoustsic and visual equipment for giving passanger information, and booking systems for seat reservation, further the central traffic control and traffic management. Railways are operated around the clock continuously, thus requiring continuous availability and the instantaneous elimination of failures.

The operating department is responsible for the transportation of goods and for passanger transportation. The maintenance department looks after track and vehicle maintenance, further the maintenance of communication and safety equipment.

Following the introduction of electric trains, the wireline networks had to be liquidated. These were at first replaced by underground cables. Later, the introduction of optical fibres constituted the fundamental building blocks of railway communication, usually implemented by a self-supporting structure over the masts of the electric power network.

The technology of transmission and switching systems applied in railway communication has always been one step before that of public telephone networks. From the onset, digital transmission has been applied for data transmission and non-

voice services such as transportation control, booking service, vehicle follow-up, realized with the most up-to-date devices. Wireless communication is also widely used in railway systems, e.g. for providing continuous communication with moving vehicles and operating marshalling yards.

Waterway communication

In high-seas navigation and river navigation, wireless communication paved the way for emergency signalling (S.O.S.) and weather reports. The use of the Global Positioning System GPS, providing globally the geographical parameters of the ship, is mandatory in most cases (e.g. for maritime sailing boats). At present, comfort services such as telephone, fax etc. are concomitant features of high-seas and river navigation.

In 1979, the United Nations Organization has established the global INMARSAT organization (International Maritime Satellite Organization), originally intended for communication over the seas, between airlines and mobile land terminals. INMARSAT is providing voice and data communication and emergency services over global oceans, excepted the polar regions.

Public highway communication

The special purpose network of public highways is applied for metropolitan public road traffic control and for emergency calls along autoroutes. Regular announcements on the public road traffic situation, primarily broadcast over radio, is gaining importance.

The wireline sub-network of metropolitan traffic control is generally separated from the public telephone network. Following the expiring of telecommunication monopolies, these subnetworks have gained importance because for recent service providers, it becomes more and more difficult and costly to acquire additional lines and to establish subnetworks. At present, these special networks and their substructure have a growing share in public communication. In emergency services, both wireline and wireless solutions are applied. The substructures of wired networks are the protecting pipelines holding fibre cables. Wireless networks are operated by antennas placed on columns or towers. The use of the GPS satellite positioning system promotes the flow of public traffic and helps the police in finding stolen

vehicles. The INMARSAT supplementary positioning system helped to improve the GPS accuracy from 100 to 5 meters.

Air traffic communication

Modern air traffic is inseparable from air traffic control, based on terrestrial and on-board radio location (RADAR) systems and the information exchange between air traffic controllers and aircraft drivers. The safety of air traffic is warranted by the co-operation of international air traffic controllers. The technological network of the air traffic is based on the application of wireless networks and the wireline communication between terrestrial traffic control centres. Air traffic is heavily related to the international booking system that is based on the special purpose network connecting airline companies. On board of aircrafts, passanger comfort services (phone, fax) are presently available but the use of mobile phones is not allowed.

4.6.2. Pipeline communication

Networks serving petroleum, gas and petroleum products

The transportation of carbon hydrogens calls for a telecommunication service having high quality, extreme reliability and availability in order to guarantee the safety of operation. These objects are normally achieved by laying telecommunication cables along the pipeline protection lane, assuring safety of operation. The main features of these special purpose networks are summarized in the following.

- Telecommunication cables are laid along the pipeline,
- Four-wire voice frequency circuits are applied for the telemechanical and dispatcher telephone services,
- Long distance communication is served by 12-channel carrier frequency equipment,
- The network is planned, established and maintained by corporations within the industry branch.

Experiences over many decades have shown that the telecommunication cables laid in the safety zone of the pipeline are much more protected than the postal cables that follow, in most cases, public roads.

Water conduit networks

The transportation of drinking water over pipelines can be traced back to ancient times. The water supply of Rome was provided in the fourth century B.C. by the first large water conduit built along the Aqua Claudia. Recently, technical solutions have been found for placing the telecommunication cable within the conduit. For this purpose, fibre transmission has been primarily applied after solving the main problem of leading out the cable from the conduit, in spite of the considerable pressure difference, by the application of special gaskets.

Drain-pipe networks

Drain-pipe telecommunication networks are not really existing. There is a possibility of placing the cable within the spout, similarly to the water conduit. Compared to the water conduit, the technical solution is simpler inasmuch as no pressure difference exists, but also more complicated because of the chemical effects endangering the cable.

District heating networks

In towns, district heating networks may also be utilized for the placement of telecommunication cables.

4.6.3. Networks for transmission of electricity

Technological requirements

The establishment of separate telecommunication systems for networks transmitting electricity is dated back to the beginning of this service. This trend was accelerated in the fifties by the development of international cooperation and the establishment of the united energy system, supported by the Union for the Coordination of Transmission of Electricity. At the beginning, this trend was confined to serve leased postal properties by applying carrier frequency transmission over power lines. Even at that time, separate two-way communication between power stations, power line substations and dispatcher headquaters was required as the production, transportation and consumption of electric power had to be continuously

controlled by the test and data processing systems and their coordinating dispatcher services.

Network services

In the nineties, the introduction of digital transmission resulted in an extremely fast development of the special purpose networks in the Central European region, in parallel with the development in Western countries. Fibre transmission and modern digital transmission techniques (SDH, ATM) and telephone exchanges provided capacities exceeding the technological requirements regarding transmission speed, transmission bandwidth and number of channels. This paved the way for the utilization of these technological systems also for business and public purposes, especially in the fields of data transmission, information services (Internet), further wireline and wiresless (GSM) transmission.

At present, carrier frequency transmission is no longer favoured though as terminal equipment, multichannel quasi-digital equipments are still in operation. Their application is confined primarily to dispatcher telephone networks and to the automation of power line protection. Due to frequency management considerations and bandwidth problems, microwave equipments are no longer competitive as compared with optical cable systems.

For applications in the power line industry, the optical network seems to be the best solution, in spite of the relatively high installation costs and the difficulties due to the simultaneous operation with electricity transmission. The optical cable installed over the power line within a protective cover has several advantages. It is completely protected against mechanical damages and electromagnetic disturbances. The network can simply be installed practically over any path, due to the densely operated power line system (120, 220, 400, 750 kV networks). The fibre capacity is practically unrestricted (nx96 fibres), the intermediate branching and amplification facilities can easily be realized. By cooperation with other service providers, a second transmission lane can be operated over the power line system cables. At lower voltage levels (20, 35,120 kV), open wire lines with plastic covering can be fastened on the power line masts. These masts have recently been utilized also as antenna towers for GSM and DCS transmissions. Further, international power line networks seem to be of interest also for trans-national service providers.

Power Line Communication (PLC)

This kind of communication, operated over the 230 V distribution network, provides local access to the customer. In the frequency range above 1 MHz, two-way data transmission with bit rates between 1 and 8 Mbit/s are provided between the customer and the PLC equipment at the last transformer station. From the transformer station onwards, data transmission is realized not over the power line network but over the conventional telecommunication network.

4.6.4. Water management network

This network is intended for protection against floods and inland waters, further for maintaining water qualilty. The protection against floods and inland waters has always been of utmost importance. With the increasing significance of environmental protection, the control of water quality has also come into prominence.

In all cases, early warning is the most important item in order to allow the authorities and the population to take timely precautions averting the danger, either by raising embankments or by evacuating the population and/or the animals, and also other economical assets.

The protection activiy, intended for saving economic goods, human lives and properties, can be classified to fall into following groups.

- Protection against floods,
- Protection against inland waters,
- Protection against frost
- Water quality damage avertion,
  - ➢ Production-type water management activities:
- Provision of water supply and canalization,
- water utilization,
- others.

In course of operating the networks for protection against floods, inland waters and quality deteroriation, radio communcation has a significant role, especially in disaster situations. In this relation, a stand-by radio service has to be operated (e.g. the TETRA system).

4.6.5. Future of technological networks

In the course of production procedures, more and more frequently information transmission is generating direct production resources, thus information networks are becoming vital elements of telecommunication networks.

The integration of public and technological networks can be witnessed, this convergence being remarkable in the field of networks and services. The double convergence has been brought about by changes in regulation (liberalization) and by technological changes (digitalization).

Liberalization allowed technological telecommunication networks, by cooperations or competitions with public networks, to provide subscriber services and thus enter into the telecommunication market. Finally, as a result of digitalization, public telecommunication networks are now capable of meeting the requirements of informatics in the field of several technological processes (e.g. interactivity).

Translated by Tamás Sárkány dr.

4.7. Broadcasting Networks

Kovács Imre dr.,author

András Gschwindt dr., reviewer

This subchapter is organized as follows. First, we introduce the program transportation networks as the distribution network to the broadcasting transmitters, by describing the possible transmission systems. Then, we review some important broadcasting systems and special solutions and finally the subsystems of the interactive video broadcasting systems are also described briefly.

4.7.1. Program Transportation Networks

The program transportation could be characterized in practice as a telecommunication service between given end-points where at least one end-point must be the play-out unit of the content provider. The play-out unit is obviously a part of the content provider's studio and ensures the transmission of the content continuously, according to the pre-defined repertoire. The content provider's studio has several tasks, such as capturing, storing and archiving video and/or audio, post-processing, news edition, etc. In this subchapter, the details of the latter mentioned systems are out of the interest, we only concentrate the used signal formats.

The types of the target end-points depend on the medium of the transmission. In the case of serving the terrestrial broadcasting the destination end-point is the television or radio trasmitter, while it is the satellite uplink as in the case of satellite broadcasting, and finally, in the case of cable or wireless distribution the destination end-point is the program distribution head-end. This structure is valid for video and audio distribution and for analogue and digital technology, but the modulations, the protocols and the distribution formats are obviously different.

A general program transportation system is shown in Figure 4.7.1. All the three important solutions are in the figure. At the left side the studios provide the source audio or video signals, the number of studios depends on the application. The network adapter adapts the incoming signals according to the applied modulation

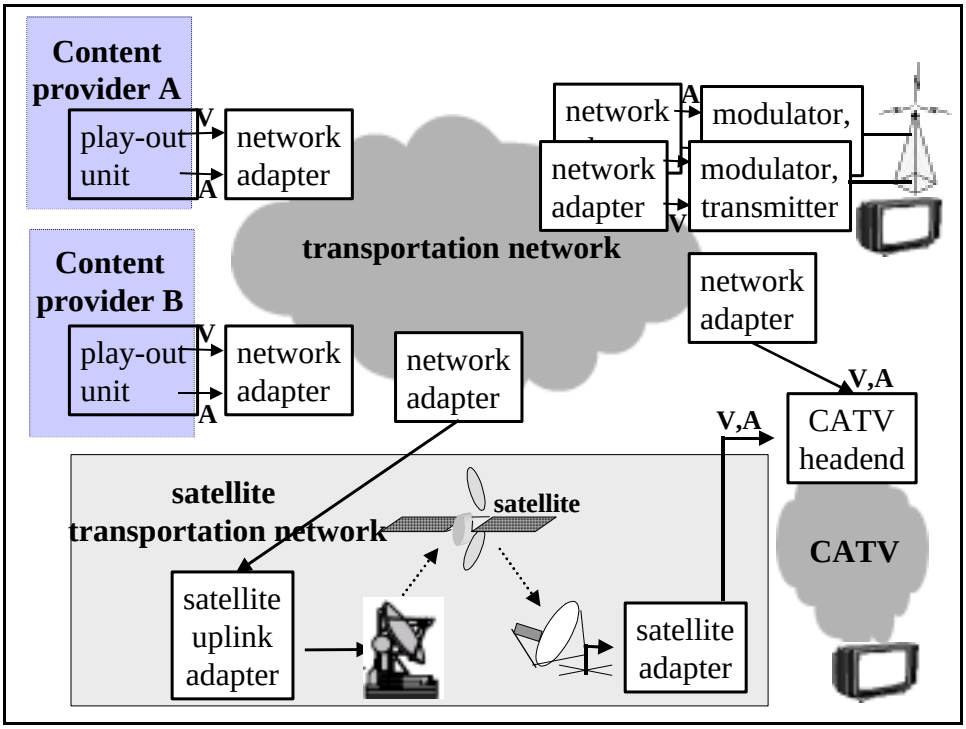


Figure 4.7.1 Scheme of program transportation networks (Term A and V stands for audio and video, respectively)

and protocol of the transportation network, hence the source format and the applied transportation network determines the network adapter precisely.

There are various implementations of the transportation network, e.g. simply broadcasting or a two-way telecommunication network whose capacity is high enough.

In the case of analogue transportation the most common solution is the FM RF micro whose transmission parameters are significantly better than those of the broadcasting standards hence the transportation does not degrade the quality of the distributed material. In this case, the source of the network adapter is a CVBS[7] FDM PAL, SECAM or NTSC [4.7.1] signal and the base-band audio channels and is modulated according to the modulation mode of the FM micro system.

In the case of digital transportation, the construction of the network adapter depends on the format of the input signal, the technology of the digital transportation network and the applied audio and video source coding/compression algorithm.

[7] Composite Video Blanking Sync

The two possible outlines of the network adapters mentioned above are sketched in Figure 4.7.2.a and Figure 4.7.2.b The construction shown in Figure 4.7.2.a. receives analogue CVBS video and analogue base-band audio and transmits compressed digital signal using appropriate source coding algorithms. Figure 7.4.2.b shows a more recent solution where the network adapter receives digital video and audio on SDI (Serial digital Interface [4.7.2], [4.7.3]). Since the bit-rate of SDI is 270 Mbit/s, the distance between the studio and the network adapter must be relatively small.

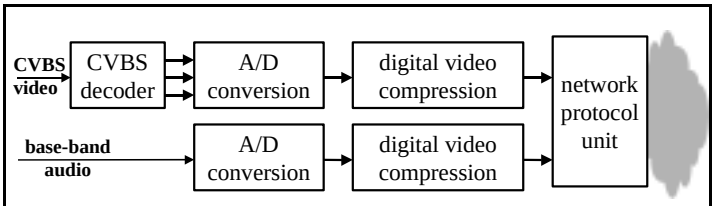


Figure 7.4.2.a. CVBS-digital network adapter

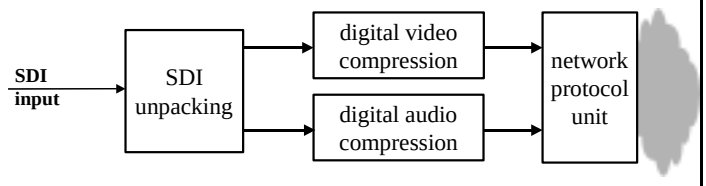


Figure 7.4.2.b. SDI-digital network adapter

The network protocol unit is responsible for the network-depend protocol encapsulation. Here we only emphasize that the bit-rate of one program could be extremely high and the on-line requirement has to be guaranteed. The on-line requirement implies that the discontinuity error and the jitter of the network must be handled inside the network or in the network adapter.

The most critical part of the chain is the video and audio compression since is determines the trade-off between quality of the decompressed material and the resulting compressed bit-rate and its cost i.e. the more the resulting bit-rate the better the quality and the higher the cost.

The first digital transportation networks used 140 Mbit/s and resulted in transparent quality after the decompression but the bit-rate and the cost of the network was extremely high hence the application of this network was limited. The next applied bit-rate was the 34 Mbits/s [4.7.4], which could be handled economically

on several networks (PDH, SDH, ATM). The distortion of the decompressed material i.e. the quality of the decoded audio and video is met the requirements but the compression is no more transparent hence further processing (editing, re-mixing) of these material causes audible distortion. For this reason this type of transportation is rarely used between studios.

The development of the compression algorithms resulted in further bit-rate reduction with almost no quality degradation in video and audio. For this reason, in case of transportation of broadcast quality video and audio much lower bit-rate is allowed (6-15 Mbit/s). The MPEG-2 video [4.7.5], MPEG-1/MPEG-2 audio [4.7.6] [4.7.7] and MPEG-2 transport stream [4.7.8] multiplexing standards are used for transportation of video and audio for some years.

4.7.2. Types of the Broadcasting Systems

The outline of the network adapters and the related instruments of analogue and digital terrestrial and satellite broadcasting is shown in Figure 4.7.3 and Figure 4.7.4, respectively.

In the case of analogue terrestrial broadcasting the network adapter does the following steps:

- receiving the input digital signal according to the protocol of the transportation network
- error checking and error correction coding
- audio and video unpacking
- de-multiplexing the audio, video and other data from the multiplexed stream since the stream may contain several channels
- decoding the de-multiplexed component by using the appropriate decoder
- encoding the audio and video according to the applied analogue standards
- modulation

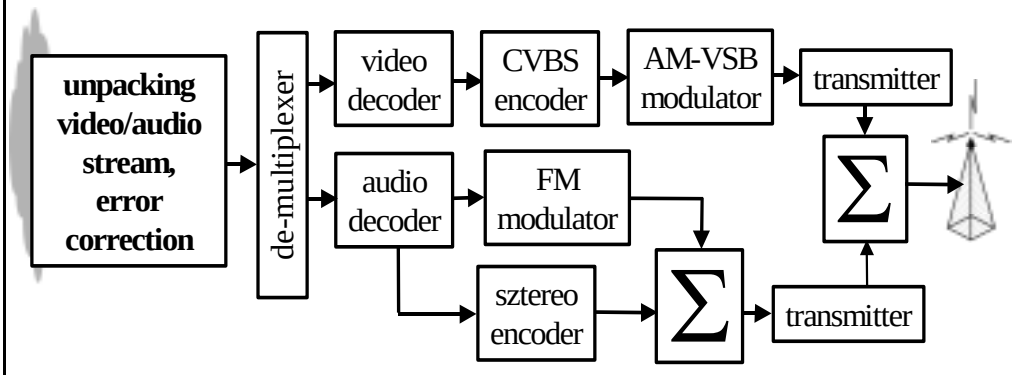
As it is shown in Figure 4.7.3.a, the analogue terrestrial broadcasting is able to broadcast two-channel stereo audio remaining compatible with mono systems by using additional coding for the second audio channel. In Europe there are two different methods for this purpose in these days. The NICAM-728 728 (Near Instantaneously Companded Audio Multiplex) [4.7.9] is a digital source coding

method, while the other system uses two sound carriers according to the recommendation ITU-R BS. 707-2 [4.7.10].

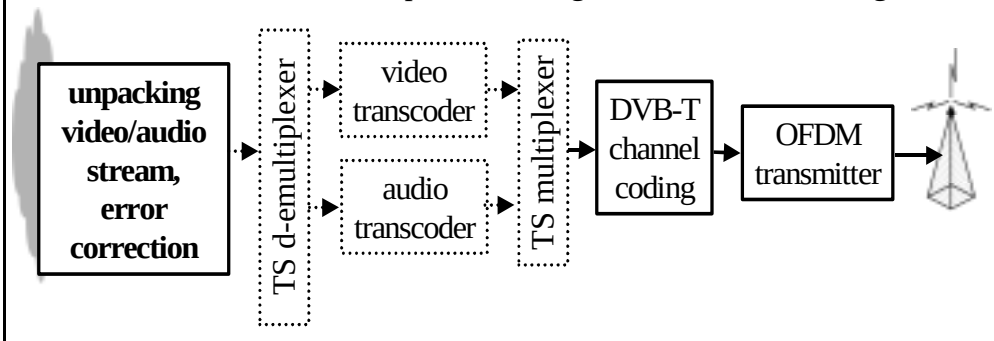
In the case of digital terrestrial broadcasting (Figure 4.7.3.b) the network adapter does the following steps:

- receiving the input digital signal according to the protocol of the transportation network
- error checking and error correction coding
- audio and video unpacking
- de-multiplexing the audio, video and other data from the multiplexed stream since the stream may contain several channels
- optional transcoding the compressed material if the source coding of the input and transmitted signal is different (different methods or same method with different parameters)
- DVB-T channel coding and OFDM modulation

The optional parts are sketched with dotted line in Figure 4.7.3.a. For instance, when the transportation network uses the ETSI 34 Mbit/s compression, the transcoders must re-encode the video and audio according to MPEG-2 video and



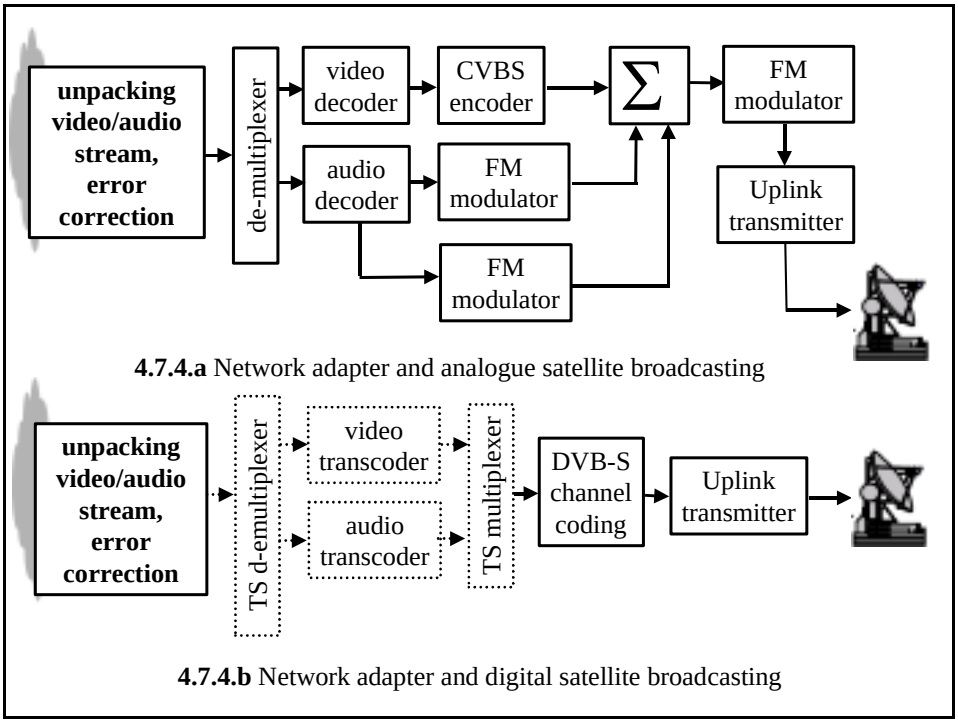
4.7.3.a Network adapter and analogue terrestrial broadcasting



4.7.3.b Network adapter and digital terrestrial broadcasting

MPEG-1 audio standards. Finally, the TS multiplexer must insert the transcoded video, audio and the ancillary data into an MPEG-2 TS according to MPEG-2/DVB-T [4.7.11]. When the transportation network also uses MPEG-2 TS format, the transcoding could be a simple bit-rate reduction re-encoding if needed. The MPEG-2 TS format allows us to broadcast several program channels and ancillary data such as EPG (Electronic Program Guide [4.7.12]) on one frequency, while the analogue broadcasting could map only one program channel into one frequency.

The feeding of the analogue and digital satellite broadcasting is sketched in Figure 4.7.4. Several parts of these systems are identical to the terrestrial broadcasting case. The TS can consist of several programs therefore the digital technology is able to transmit multiple programs using one transponder.



4.7.4.a Network adapter and analogue satellite broadcasting

4.7.4.b Network adapter and digital satellite broadcasting

In contrast to digital satellite broadcasting, the analogue one can only transmit one program (one video channel and several audio channels) on one transponder hence the de-multiplexer of the network adapter must extract the transmitted program from the transportation network. The analogue composite video signal is modulated using FM, while the audio channel is modulated using a noise reduction technique and FM. The ancillary audio channels could be multiplexed using several sub-carriers and FM modulation, or another prevalent technique is the digital modulation and digital compression of the ancillary audio channels, where the source coding is usually a variant of MPEG-audio.

The feeding of the digital satellite broadcasting transponder is similar to the terrestrial case, the only difference is the channel modulation, which is DVB-S (DVB-Satellite [4.7.13]) which is discussed later.

4.7.3. Digital Video Broadcasting

From the eighties the digital technology allows of the digital transportation of the programs, but the required native (uncompressed) bit-rate could be extremely high hence from the end of the eighties to the beginning of nineties several compression methods are developed to use up the bandwidth economically.

The several incompatible compression methods could not co-operate hence the ISO/IEC established the working group ISO/IEC/JTC1/SC2/WG11, and the abbreviation of this working group was ISO/MPEG (Moving Pictures Experts Group).

Beside this, the broadcast providers and consumer electronic companies proposed a common pan-European platform for digital broadcasting in 1991, and at the end of that year the ELG (European Launching Group) committee is established to follow up the development of the European digital broadcasting. In 1992, the ELG published the Memorandum of Understanding, which contains considerations for the developers. The committee established a new organization, the DVB (Digital Video Broadcasting) project whose main purpose was the marketable and harmonized development of the cable, terrestrial and satellite digital television.

The DVB project has chosen the MPEG-2 system, video and audio standard which defines efficient compression, scalable coding quality and reliable data structure.

Currently, the DVB project has developed the transmission method of every important medium. The channel coding has been developed according to the property of the medium and the main purposes of the transmission, while the robustness was the most significant consideration.

In the following, we briefly outline the MPEG-2 data structure which plays an important role in DVB.

The MPEG-2 standard

The MPEG-2 standard consists of several parts. The first part (ISO/IEC 13818-1 [4.7.5]) defines the system i.e. the syntax and semantics of an MPEG-2 stream. The video and audio coding is defined in the second and third part of the standard i.e. in ISO/IEC 13818-2 [4.7.5] and ISO/IEC 13818-3 [4.7.7], respectively.

The output streams of the video and audio encoder are called video and audio elementary streams (ES). The basic elementary stream could be packetizes (PES – Packetized ES). The first part of the standard (MPEG-2 system) defines how video, audio and other elementary streams are multiplexed into one stream. The MPEG-2 system defines some important ancillary information which required of synchronized decoding. The synchronized decoding is carried out by using time stamps. The synchronized decoding is essential to present the video and audio signal in time. A PES contains one program component (video, audio, teletext, etc.) hence one program consists of several PES-s and these PES-s must be multiplexed into one stream.

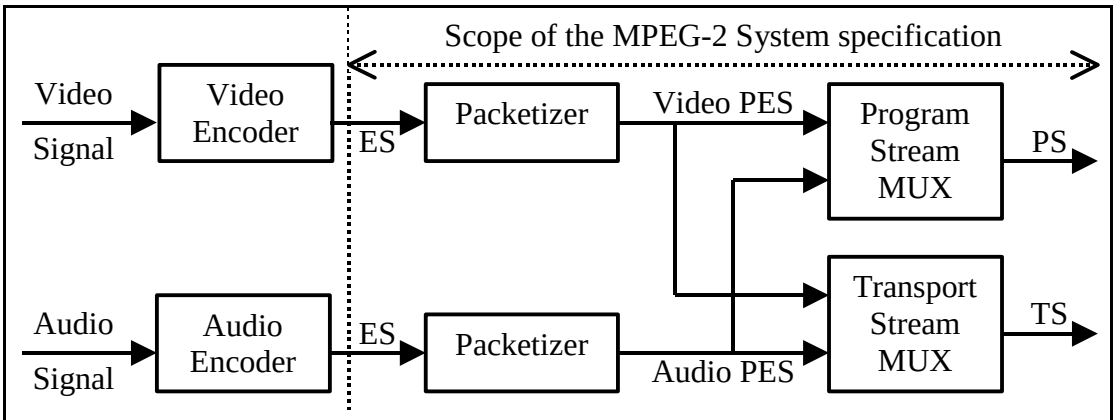


Figure 4.7.5. The generation of MPEG-2 PES, PS and TS

The MPEG-2 system defines two multiplex methods, the transport stream (TS) and program stream (PS). TS is optimized for noisy channel hence the this type of multiplexing could be easily adapted for co-operating with error detection and correction algorithms while the PS is developed for error-free transmission or storage. Since the broadcast channels are noisy channels, only the TS is used for broadcasting. The relation of PES, TS and PS is shown in Figure 4.7.5.

As we mentioned above, the TS is able to carry several independent programs in one stream. Furthermore, the TS can multiplex the programs with independent

time reference, which is important when the video and audio encoders of different programs have definitely different time references hence the time stamps and other synchronization informations must be different. This nice feature of TS allows the content providers to encode their own material anywhere, usually in the place of the production of the material.

The ancillary data of the programs in an MPEG-2 TS are stored in a separate structure called Program Service Information (PSI). The content of the PSI is transmitted using special tables.

Transmission Modes and Systems Defined by the DVB Project

The following transmission modes are developed by the DVB project:

- DVB-S (Satellite) [4.7.13],
- DVB-C (Cable) [4.7.22],
- DVB-T (Terrestrial) [4.7.11],
- DVB-CS (Community System) [4.7.14],
- DVB-MS (Multi-point Satellite) [4.7.15],
- DVB-MC (Multi-point Cable) [4.7.16].

The DVB standard contains further transmission modes and several important parts, but they are not discussed here due to the lack of extent.

The different DVB transmission modes contain similar parts as much as possible, and it is the same from the viewpoint of transmission what type of data (audio, video, binary data) are transmitted. The same parts of the different transmission modes could be implemented on one chip. In practice, the demodulators of the three different channels and the same decoder of a DVB receiver could be implemented on one chip.

DVB-Satellite

The DVB-S is a broadcasting and transportation system in the 11/12 GHz Ku-band. The DVB-S standard [4.7.13] defines the modulation and channel specification of the digital satellite multi-program normal and high-definition broadcasting. The outline of the DVB-S encoder is shown in Figure 4.7.6.

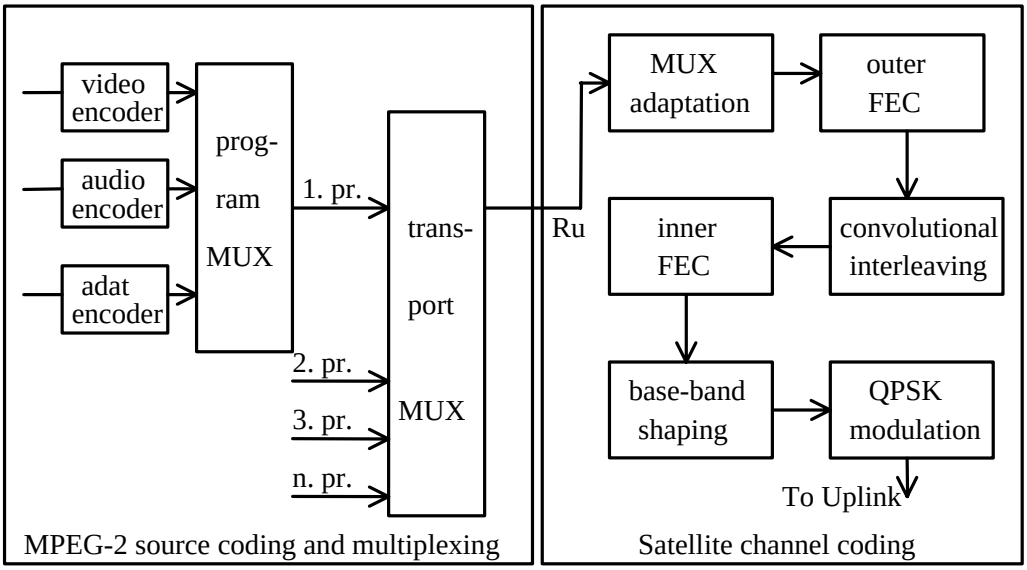


Figure 4.7.6. The generation of DVB-S signal

The main development parameters of the modulation was the following:

- Equivalent Isotropically Radiated Power (EIRP) of the satellite transponder
- non-linear distortion of the transmitter
- open-air attenuation computed from the distance (205 dB)
- attenuation of the atmosphere
- figure of merit (gain/temperature: G/T)

The received power is the limiting parameter of the receiving hence the sufficient robustness against noise was the most important requirement. For this purpose, the DVB-S uses both inner and outer error correction coding.

The non-linear distortion of the transponder and the strong attenuation of the atmosphere require such a modulation mode that provides acceptable bit-error rate even if the signal-to-noise ratio is very low. These facts determine the modulation, which could be no other than QPSK (Quarternary Phase Shift Keying).

To guarantee the required bit-error rate, the DVB-S use high-efficient error correction algorithm. The error correction algorithm consists of punctuated convolution coding and Reed-Solomon block coding for error correction and use convolution interleaving to spread out the error burst. The overall algorithm has both bit and byte error correction capability, and long error burst periods could be also detected and corrected up to a limit. Furthermore, the convolution coding allows high-efficient transponder bandwidth adaptation.

Although the overall system was designed for optimal TDM transmission where several programs are transmitted using only one carrier (MCPC: Multi Channel Per Carrier), it is possible to use multiple carrier FDM where one carrier is modulated by only one program (SCPC: Single Channel Per Carrier).

The applied error correction coding ensures that the transmission is quasi error free (QEF: at most one uncorrected error in one hour) when the carrier-to-noise ratio is above the threshold.

DVB-Terrestrial

The DVB-T standard [4.7.12] defines the modulation and channel specification of the digital terrestrial multi-program normal and high-definition broadcasting. The main requirements of the development of the DVB-T standard was the following:

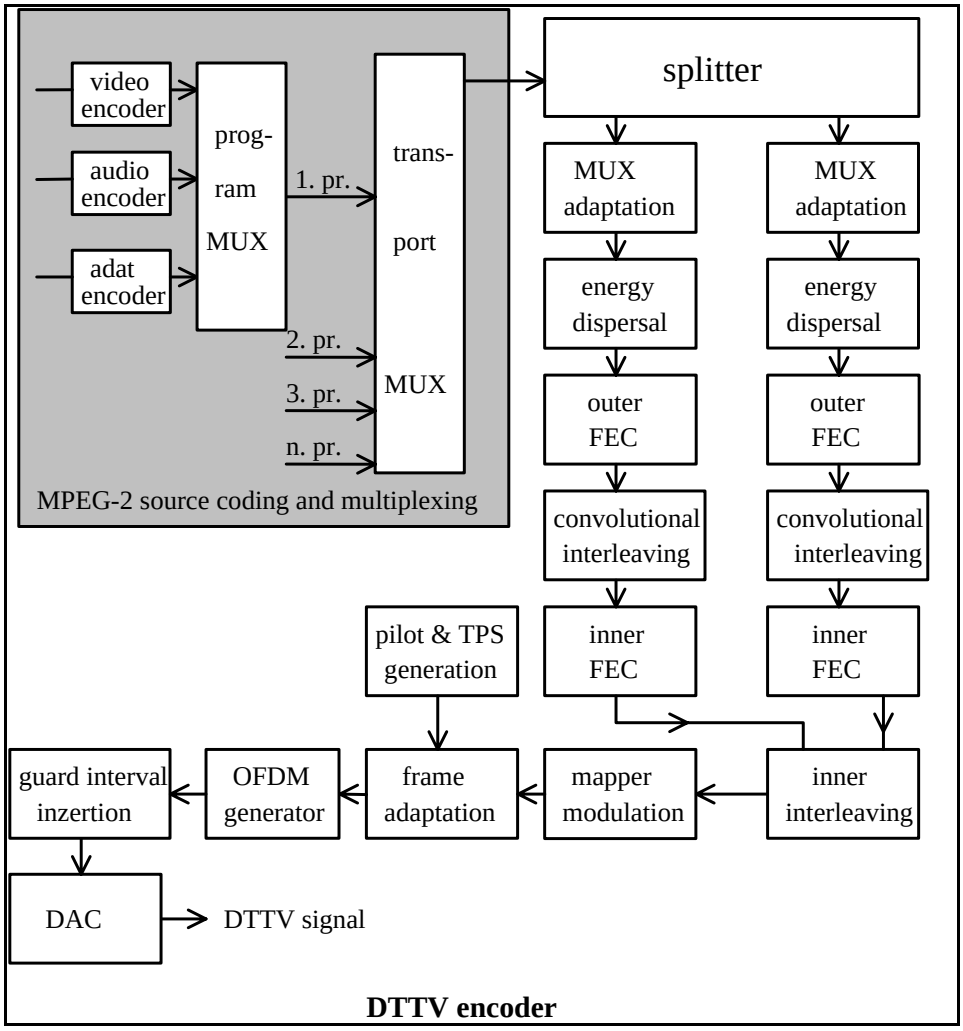
- the emitted signal must fit into the existing 8 MHz UHF channel
- robustness against noise inside one channel and interference
- spectrum efficiency: SFN networks
- capability for mobile and in-door receiving (OFDM: Orthogonal Frequency Division Multiplex)

The OFDM supplemented with error corrections and interleaving ensures efficient noise protection, which allows both mobile and metropolitan area receiving. Similarly to DVB-S, the applied error correction coding ensures that the transmission is QEF when the carrier-to-noise ratio is above the threshold.

The outline of the DVB-T encoder is shown in Figure 4.7.7.

4.7.4. DVB Interactive Systems

During the development of the DVB systems the need of the interactive multimedia systems had been also occurred hence the DVB-T project established the SIS (System for Interactive Services) working group, which had to develop the system model of the interactive services. The system was standardized in 1997. The most important property of the system model is the high capacity download channel (from the service provider to the user) while the return channel (from the user to the service provider) has significantly lower throughput. While the bandwidth of the



4.7.7. DVB-T channel coding

download channel could be even the bandwidth of the broadcasting channel i.e. 48 MBit/s, the maximal throughput of the return channel is only 150 kbit/s.

The system model allows that the broadcast service provider (BSP) and the interactive service provider (ISP) could be different, the only requirement is that they must have an interactive connection with enough high bandwidth. The service contains digital television service supplemented with interactive services such as NVOD (Near Video On Demand), e-mail, web, on-line shopping, etc. The system model of the SIS working group consists of the following layers:

- Physical layer: channel coding, modulation, etc.
- Transport layer: data structures and communication protocols
- Application layer: interactive software and its environment

This system model is a simplified OSI model in practice. The reason for the simplification was to facilitate the implementation. The following channels are definen in the SIS system model:

- Broadcast channel: one-direction channel containing video, audio and data
- Interactive channels: two channels between the user and the service provider

  ➢ Return channel: small capacity channel form the user to the service provider carrying user requests and user responses

  ➢ Download channel: from the service provider to the user used for download the requested data, but this channel could be encapsulated in the broadcast channel

The specifications of the physical and transport layers are defined in the standard of the applied transmission mode (DVB-S, DVB-T). There are several possibilities for interactive channels in DVB, for instance:

- PSTN/ISDN [4.7.18],
- DECT [4.7.19],
- GSM [4.7.20],
- Satellite [4.7.21]
- etc.

Finally, the general construction of the DVB interactive system is shown in Figure 4.7.8.

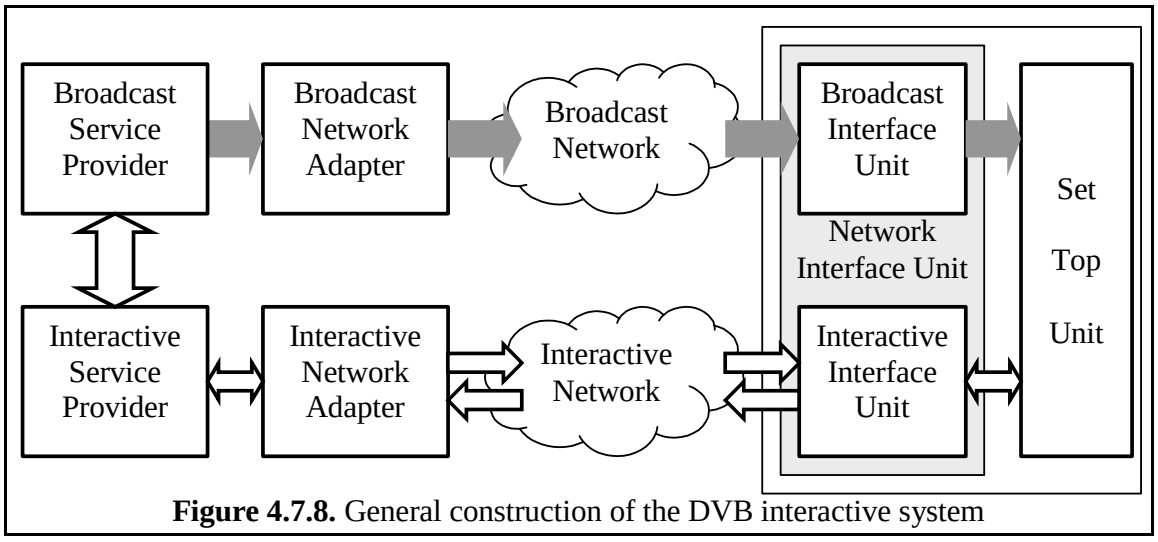


Figure 4.7.8. General construction of the DVB interactive system

References

[4.7.1] Recommendation ITU-R BT.470-4 Conventional, Enhanced And High-Definition Television Systems;

[4.7.2] Recommendation ITU-R BT.656-3, Interfaces For Digital Component Video Signals In 525-line And 625-line Television Systems Operating At The 4:2:2 Level Of Recommendation ITU-R BT.601 (Part A);

[4.7.3] Recommendation ITU-R BT.672-3, Video-Frequency Characteristics Of A Television System To Be Used For The International Exchange Of Programmes Between Countries That Have Adopted 625-line Colour Or Monochrome Systems;

[4.7.4] ETSI ETS 300 174/A1, Network Aspects Digitalis Coding of Component Television Signals for Contribution Quality Applications In The Range 34-45 Mbit/s;

[4.7.5] ISO/IEC 13818-2, Generic Coding of Moving Pictures And Associated Audio: Video;

[4.7.6] ISO/IEC 11172-3, Coding of Moving Pictures And Associated Audio For Digital Storage Media At Up To About 1.5 Mbps – Part 3 Audio;

[4.7.7] ISO/IEC 13818-3, Generic Coding of Moving Pictures and Associated Audio: Audio;

[4.7.8] ISO/IEC 13818-1, Generic Coding of Moving Pictures and Associated Audio: Systems;

[4.7.9] ETSI EN 300 163, Television Systems; NICAM-728: Transmission System Of Two-Channel Digital Sound With Terrestrial Television Systems B, G, H, I, K1 and L;

[4.7.10] Recommendation ITU-R BS. 707-2, Broadcasting Service (Sound), Transmission Of Multisound In terrestrial Television Systems PAL B, G, H And I, And SECAM L;

[4.7.11] ETSI EN 300 744 (1999-07), Digital Video Broadcasting (DVB); Framing Structure, Channel Coding And Modulation For Digital Terrestrial Television;

[4.7.12] ETSI ETS 300 707 (1997), Electronic Programme Guide (EPG); Protocol For A TV Guide Using Electronic Data Transmission;

[4.7.13] EN 300 421 (1997-08), Digital Video Broadcasting (DVB); Framing Structure, Channel Coding And Modulation For 11/12 GHz Satellite Services;

[4.7.14] EN 300 473 V1 (1997-08), Digital Video Broadcasting (DVB); Satellite Master Antenna Television (SMATV);

[4.7.15] EN 300 749 (1997-08), Digital Video Broadcasting (DVB); Microwave Multipoint Distribution Systems (MMDS) Below 10 GHz;

[4.7.16] EN 300 748 (1997-08), Digital Video Broadcasting (DVB); Multipoint Video Distribution Systems (MVDS) At 10 GHz And Above;

[4.7.17] ETS 300 802 (1999-11), Digital Video Broadcasting (DVB); Network-Independent Protocols for DVB Interactive Services;

[4.7.18] ETS 300 801 (1997-08), Digital Video Broadcasting (DVB); Interaction Channel Through Public Switched Telecommunications Network (PSTN) / Integrated Services Digital Networks (ISDN)

[4.7.19] EN 301 193 (1998-07), Digital Video Broadcasting (DVB); Interaction channel Through The Digital Enhanced Cordless Telecommunications (DECT)

[4.7.20] EN 301 195 (1999-02), Digital Video Broadcasting (DVB); Interaction Channel Through The Global System for Mobile Communications (GSM)

[4.7.21] EN 301 790 (2000-12), Digital Video Broadcasting (DVB); Interaction Channel For Satellite Distribution Systems;

[4.7.22] EN 300 429 (1998-04), Framing Structure, Channel Coding And Modulation For Cable Systems;

Web links:

www.etsi.org

www.itu.org

www.dvb.org

www.digitag.org

www.drm.org

www.mpeg.org

4.8. Cabled Distribution Networks for Television and Sound Signals (Broadband Cable Networks)

Gábor Mátay dr., author

Sándor Stefler, reviewer

4.8.1. From Community Antenna Television Systems to Broadband Cable Networks

The history of program distribution on cable began long time ago and it started with the distribution of audio programs for subscribers. Hungarian examples are: the so-called “News on Telephone” (Telefonhírmondó in Hungarian) service and wire broadcasting system (WBS). The WBS was operated in the 1950s. The mass spreading of television (TV) sets, the increasing number of available TV channels for reception in a given place and the building of the new block of flats needed the application of community reception instead of individual one. The introduction of community reception for TV channels was advantageous from points of view of the economic efficiency, look the town and EMC[8]. With the introduction of the community antenna television systems “forests of antennas” disappeared on the roofs of buildings and the occurred mutual interference in consequence of oscillator signals and their harmonics from TV sets radiated by individual receiving antennas was stopped.

The community antenna TV system is a one-way system, which serves for the local cabled distribution of radio and TV channels received by community antenna. The system consists of head-end, cabled distribution network and subscribers’ outlets. Depending on the number of distributed TV channels and the network dimension it has two types: a small (main antenna TV system) and a large one (community antenna TV system). General characteristic of both types is *one-way information transmission*. In case of community antenna TV systems air- or ground cables transmit the signals. Wide-band amplifiers are inserted into the different parts

[8] EMC - **e**lectro**m**agnetic **c**ompatibility

of distribution network in order to compensate the attenuation of cables and other passive elements. Cable television (CATV[9]) systems operated all over the world at present developed from community antenna TV networks, but these were changed into qualitatively new systems [4.8.1].

The basic difference between *CATV networks* - nowadays they are often called *broadband cable networks* - and community antenna TV networks is that the CATV systems distribute more TV channels (the number of TV channels increased significantly in consequence of program distribution by DBS) and the CATV systems are *interactive*. The interactivity means that the CATV system by using its return- and forward path is able to receive commands or data from subscribers and react to them. The interactivity needs *two-way information transmission* from the head-end to the subscribers. So the CATV systems have two-way distribution networks, which afford possibility to introduce new so called *value added services* connected to program distribution and/or data transmission (see under 5 and 6 chapters especially under 5.5, 6.2, 6.3, 6.4 and 6.9 subsections). Nowadays these new services get more importance beside the function of program distribution.

The present change and future development in broadband cable networks (CATV systems) are determined by following things:

- *the convergence* among telecommunication, broadcasting and multimedia services;
- the *development of services* on CATV systems *in the fields of program distribution* (PPV[10], interactive TV[11], VOD[12], NVOD[13]) *and data transmission* (Internet, telephony, teleshopping, telebanking, data collection of public works /reading gas-, electric meter, etc./, teleworking, telemedicine, etc.);
- the development in the technique of signal processing and transmission, *the quick spread of digital systems in the audio- and video broadcasting;*

[9] CATV - **Ca**ble **T**ele**v**ision

[10] PPV: **p**ay-**p**er-**v**iew programming is sold on a per program-basis, but the subscriber must order the wanted program in advance at the operator or at an automated system.

[11] Interactive TV in a wider sense means that the televiewer can take part in the watched TV program or able to influence the events in the program.

[12] VOD - **v**ideo **o**n **d**emand systems allow the subscriber to select what he or she wants to see, when he or she wants to see it. Programs are stored on a video file server. Advanced systems include so-called VCR functions of pause, rewind, and fast-forward for the subscribers at any time. Naturally the price of this service bases on the used time and the number of demanded programs.

[13] NVOD - **n**ear VOD is a service in which a predetermined program (for example selected movie) is repeated sequentially according to a determined frequency. The subscribers apply for watching a program in an adequate time for them and in case of its acceptance they get permission for access.

- the technological development in the transmission mediums and equipments;
- *integration of small and medium CATV systems into large system* in consequence of the quick growth of services and the concentration of capital.

The above-mentioned facts influence each other separately and mutually and in consequence of this interaction multifarious broadband, interactive CATV systems with economic operation are built. With the integration small and medium networks into large systems and widening data-type services *the importance of network management increases.* The technological development and the quick spread of digital systems in the audio- and video broadcasting (DAB[14], DVB[15]) play prominently important role.

In the field of DVB the most important European decision was *to apply MPEG-2[16] standards for source encoding in audio-, video- and system level*. The DVB standards (DVB-S for satellite, DVB-T for terrestrial broadcast, DVB-C for transmission on cable) worked out by ETSI[17] [4.8.2] *use MPEG-2 for video compression* and *MPEG-2 layer-II for audio compression*.

This yields several advantages, namely:

- more TV program can be put into the bandwidth corresponding to one analogue TV channel (for example satellite transponders using DVB are able to transmit 6-8 times more of TV programs than in case of analogue transmission);
- the DVB used complete digitalization opens the way toward the world of electronic program guides (EPG), Internet, data transmission, interactive TV in higher level, etc..
- the way of information transport (packetized elements of the transport stream) in case of DVB permits of the quasi error free data transmission for all kind of data through different type of transmission media (satellite transponder, terrestrial broadcast, CATV network, MMDS[18]);
- the DVB transmission system is transparent for SDTV[19], HDTV[20], sound of arbitrary quality and any kind of data.

[14] DAB - **d**igital **a**udio **b**roadcasting

[15] DVB - **d**igital **v**ideo **b**roadcasting

[16] MPEG-2 : a new version of the standard for video-compression worked out by "**m**oving **p**icture **e**xpert **g**roup".

[17] ETSI - the **E**uropean **T**elecommunications **S**tandards Institute

[18] MMDS - **m**icrowave **m**ultipoint **d**istribution **s**ystem.

[19] SDTV - **s**tandard **d**efinition **t**ele**v**ision is the collective name of 12 video formats for MPEG TV. These are lower resolution video formats of 18 video formats, which are summarized in the standard

Development of CATV technology based on the application new transmission media. Besides *the coaxial cable* (at first this was used exclusively as transmission media in CATV networks) came up *the optical cable* and lately in the last section of large distribution network especially in case of not densely populated areas *the radio channel* (MMDS) was preferred. On the basis of economic consideration one distribution network uses usually multifarious transmission media. At present the most of modern CATV network has HFC[21] technology. In consequence of continuous development the end point of optical section in CATV distribution network gets closer to subscriber's outlet.

The broadband cable networks use three kinds of *multiplexing procedures* for the transmission of TV, radio programs and data on common trace at the same time. There are *frequency division multiplexing* (FDM), *space division multiplexing* (SDM) and *time division multiplexing* (TDM) systems. It is possible that different multiplexing procedures are used jointly in one network. The most of CATV systems operated presently use frequency multiplexing procedures. The application of code division multiplexing (CDM) procedure in CATV networks is expected in the far future.

4.8.2. Building up of the CATV Systems

The place of CATV network in the program transmission *reference network* is shown in Figure 4.8.1. The parts of reference network are: *international network*, *regional network* and the national network. Inside of latest one are the broadcasting network and *CATV network (head-end and distribution network).* International and regional parts are only in extraordinarily large systems.

CATV systems consist of head-end, distribution network and at its end house networks and subscribers' outlets.

---

worked out for digital television set by the ATSC (Advanced Television Standards Committee in the USA). SDTV video formats encompass various combinations of 16:9 and 4:3 aspect ratios and frame rates of 24, 30 and 60 frames per second [4.8.3].

[20] HDTV - **h**igh **d**efinition **t**elevision accounts for six of the 18 video formats established as part of the standards for digital television set by the Advanced Television Standards Committee or the ATSC. Of the 18 formats, these are the highest resolution formats. All six formats feature a 16:9 aspect ratio and accommodate frame rates of 24, 30 and 60 frames per second. [4.8.4].

[21] HFC - **h**ybrid **f**iber **c**oaxial.

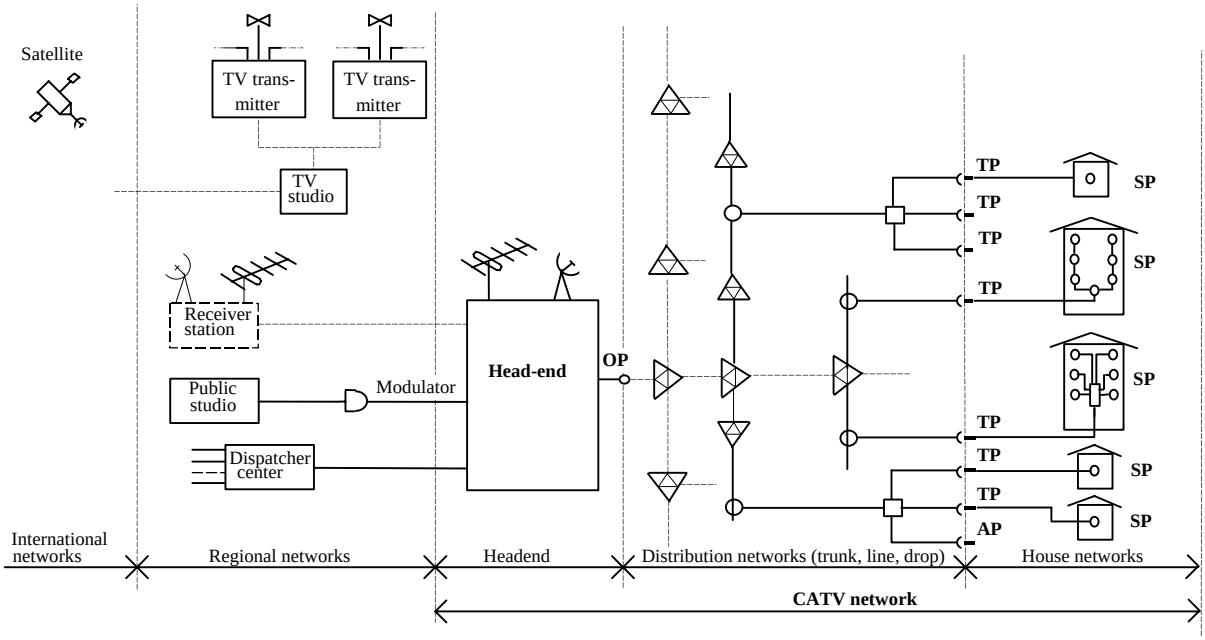


Figure 4.8.1 The place of CATV network in the reference network

The head-end converts the signals arrived from receiver station, and/or antennas, as well as other sources; and transmits them in suitable form toward distribution network. The antennas also belong to the head-end if there is not receiver-station. Depending on the state of development of CATV systems the tasks of head-end are following:

- receiving centre for terrestrial-, satellite- analogue and digital programs;
- reception point for the signals from receiver station and programs made in local studio;
- connecting point of data transmission services;
- centre of surveillance and data processing connected to the system operation.

Building up of the head-end depends on the applied multiplexing procedure, the services afforded by the CATV system, the size of distribution network and also subscribers' receivers. In case of FDM the system technique building up of the head-end is described in 2.2.3 chapter.

The *distribution network* of CATV system consists of the following parts: *trunk-*, *line-* and *drop network*.

The trunk is a portion of distribution network which is situated between head-end and line network. The purpose of the trunk is to transport the origination signals (multiplexed by head-end) at long distance to the line-networks joined with major subscriber groups.

In the CATV systems with mixed technology, in case of *so-called HFC distribution networks* the trunk uses optical cable for transmitting signals. Optical trunk consists of electrical-optical (E/O) converters (optical transmitters), monomode optical cable and optical-electrical (O/E) converters (optical receivers). The *O/E converter at the end of optical trunk* has more RF outputs (its RF amplifier is so-called bridger). The name of this end point is *optical node* (ON[22]), to which the line network or drop network (if line network does not exist) is connected. In the optical part of distribution network separated optical fibres are usually used for the transmission of signals in forward and return path. The FDM signals are transmitted with intensity modulation on the optical cable. The attenuation factors of monomode optical fibre at the wavelength of 1310 nm are 0.4 dB/km, and 0.25 dB/km at 1550 nm. Usually in the CATV distribution network the wavelength of 1310 nm is used on the basis of economic considerations (laser diode for optical transmitter is cheaper at this wavelength). In the temperature range occurred in CATV practice the attenuation of optical fibre is approximately constant with temperature and is independent of RF frequency (see 2.1.4 chapter).

Distribution networks used only coaxial cable (nowadays their occasions is very rarely) build up of coaxial cables with very small attenuation factors and broadband amplifiers of high linearity. The output and input coupling of signals is accomplished at bridger amplifiers. The frequency and temperature dependence of the cable attenuation are equalized by controlled RF amplifiers (see in 2.2.3 chapter). This is necessary because the linear temperature coefficient of cable attenuation is relatively large, it is about 0.2 %/K.

The *line-network* is a part of CATV distribution network between trunk and drop network. In forward path it distributes signals from trunk for smaller subscribers' groups, respectively in return path it collects the signals originated from subscribers for the trunk. In the modern CATV systems the line network has coaxial technology. The coaxial line-network consists of coaxial cable with small attenuation, controlled or uncontrolled RF amplifiers, program distribution filters[23] and power dividers or taps.

[22] ON - **o**ptical **n**ode, or instead of it is often used ONU for the abbreviation of **o**ptical **n**ode **u**nit expression.

[23] Program distribution filters: These filters makes possible the reception or exclusion of program packages ordered to predetermined prices in CATV distribution networks of FDM system.

The *drop network* is a part of distribution network between tap and the input of house network. Usually it does not contain amplifier, only taps and distribution filters.

The *house network* connects to CATV distribution network through transfer point (TP). This network is inside of one building or group of buildings and it ends subscribers' outlets (it is signed SP in Figure 4.8.1). The task of house network is the distribution of forward signals and the collection of return signals. The distribution of signals is possible in accessible way for every subscriber (basic services) and selectively, addressed to certain subscriber (additional services for extra subscription rate).

The structure of distribution network may be *tree-branch, ring, star* and *mixed topology*. Most of the CATV systems with a lot of TV programs and other program- and data-transmission services have usually *mixed topology*. The different kinds of topologies may occur separately and in mixed way also in the trunk-, line-, drop- and house networks. Although the ring topology is applied mainly in trunk network for increasing the reliability of services; the star structure is usually used in house networks (in FDM systems the reception of additional programs for extra subscription rate may be disabled by inserting band stop filters; that is the not paying subscribers may very easily disconnect without any change in the services for the others). The choice of topology is determined by the arrangement of subscribers in the built-up area, economic consideration (finding the minimum of the total length of cable used in the distribution network) and requirements for the reliability of services.

4.8.3. System Performance of CATV Networks

For short reviewing system parameters of CATV networks are assumed that the system use FDM, its distribution network has HFC technology, and subscribers' TV sets able to receive only AM-VSB[24] analogue TV signals. This last assumption means that these TV sets need additional equipment, so-called STB[25] for the reception of DVB-programs. In this case the STB is a digital receiver with QAM demodulator, MPEG decoder and some sort of descramblers, which make the

[24] AM-VSB - **v**estigial **s**ide **b**and AM

[25] STB - **s**et **t**op **b**ox is an additional equipment put on the top of TV set, which makes the reception of TV and radio programs transported on CATV distribution network possible. These programs cannot be received without STB.

reception of scrambled programs possible. Usually the STB is connected to the TV set in baseband that is to the video- and audio input of TV set. In case of the reception radio programs audio output of the STB is connected to the input of stereo amplifier.

Frequencies used in CATV system may be classified outer- and inner-frequency ranges. The sum of frequency bands used by programs took into the system is called *the outer frequency range of the system*. The sum of the frequencies generated in head-end and frequency bands used by program channels and data transmission is *the inner frequency range of the system*.

In order to provide for two-way stream of information in the distribution network the whole frequency range of CATV system (from 5 MHz up to 862 MHz) have to split for the forward and return path. Frequency band from 47 MHz up to 862 MHz is used for program distribution (nowadays the part from 47 MHz up to 124 MHz is rarely used because this band is applied for other purpose). Inside of this band is placed the FM-radio band from 87.5 MHz up to 108 MHz as well as the forward data transmission band. The bandwidth of one analogue TV channel is 7 MHz in VHF band and 8 MHz in UHF band. The analogue TV and radio programs in the RF channels of satellite transponder frequency band from 11.7 GHz up to 12.75 GHz are transmitted with FM modulation; but the digital TV and radio programs arrive to the receiver antenna of head-end with QPSK[26] modulation. The LNC[27] converts the received signals into frequency band from 950 MHz up to 2150 MHz then the signals get into the input of analogue or digital channel converters of head-end. In the channel converters frequency and modulation conversion is accomplished. The output signals of channel converters are combined. In this way obtained FDM channels get into the distribution network. The analogue TV programs are distributed with AM-VSB modulation in TV channels in the frequency range up to 630 MHz. Digital programs in one RF channel of satellite transponder (usually 8 TV- and 8 radio programs) get from the digital unit of head-end with m-QAM[28] modulation into

[26] QPSK - **q**uadrature **p**hase **s**hift **k**eying

[27] LNC - **l**ow **n**oise **c**onverter, it is built into the satellite receiver antenna.

[28] QAM - **q**uadrature **a**mplitudo **m**odulation: m-QAM has m-modulation state in constellation diagram, m = 16, 64, 128, 256 (note: the QPSK corresponds functionally to 4-QAM.

channels of 8 MHz bandwidth in the frequency range from 630 MHz up to 862 MHz of the distribution network.

On the basis of data sheets of firms manufacturing CATV building elements according to the requirements of system managers the usual frequency band splits are following:

- lowsplit system with subsplit): from 5 (3, 10) MHz up to 25 (26, 30) MHz for the return path and from 47 (45) MHz up to 862 MHz for the forward path, or from 5 MHz up to 30 (33) MHz for the return path and from 54 (50) MHz up to 862 MHz for the forward path;
- lowsplit system: from 5 MHz up to 55 (50, 68) MHz for the return path and from 70 (87) MHz up to 862 MHz for the forward path;
- midsplit system: from 5 MHz up to 108 (110) MHz for the return path and from 174 (150, 164) MHz up to 862 MHz for the forward path;
- -highsplit system: from 5 MHz up to 168 (174, 180) MHz for the return path and from 230 (222, 225 260) MHz up to 862 MHz for the forward path.

In Europe, so in Hungary also the lowspit systems are widespread. In forward path for distribution of analogue program channels the frequency range up to 630 MHz and for digital program channels the band from 630 MHz up to 862 MHz are used. At the same time this fact means that CATV systems have mixed: *analogue-digital techniques* in nowadays and in the near future. This temporary status of mixed techniques in CATV systems will exist until operation of analogue terrestrial broadcasting because we can hope the widespread of digital TV receivers only from this time. The distribution of digital programs in the range at the end of forward frequency band can be explained by the fact, that their transmission is possible at lower signal-to-noise ratio of 20 dB then in case of analogue programs. Attenuation of coaxial cables is higher at the upper end of forward frequency band (their attenuation is proportional to $\sqrt{f}$, see in 2.1.3 chapter).

CATV system in the reference network is between regional network and subscribers' outlets (they are signed with SP in Figure 4.8.1). *Technical parameters describing quality of whole system* are applied to signal transmission way between input of head-end (or receiver station) and subscribers' outlets.

The field strength generated by transmitted programs at receiving antenna or modulated RF signals carry the information to be distributed at the input of CATV system. In the course of the distribution of this information through CATV system is

distorted in consequence of linear and nonlinear distortions as well as reflections, and noises and disturbing signals added to it.

The level of signals and quality carrying information and the level of noises and interfering signals at the outputs of CATV system (at subscribers' outlets /SP points in Figure 4.8.1/) has to meet the technical specifications. This is checked by the test of video and audio signals at the output "converters" (TV picture tube, loudspeaker) of a so-called *reference TV set*. It is important to note in connection with the levels of signals, noises, intermodulation (discrete and composite /CTB[29] and CSO[30]/) and crossmodulation products, that the *nominal impedance at all connection points of CATV system is 75 ohms* (this equals to the value of characteristic impedance of coaxial cables used in the distribution network). The levels of noises and nonlinear distortion products at different points of the system are usually specified in dB related to carrier level. Travelling from the output of head-end (it is signed with OP /**o**riginator **p**oint of distribution network/ in Figure 4.8.1) toward subscribers' outlets the level of noises and distortion products increases. At transmission of program channels of fixed number the dimension of distribution network is restricted by the requirement that a minimum values of signal-to-noise ratio and signal-to-distortion ratio are needed at the subscribers' outlets. In other words this means that in case of a given system dimension the number of transmissible channels is determined mainly by the nonlinear distortions not taking into account the usable frequency band for program distribution [4.8.5].

In case of distribution of analogue TV channels *the reflections* in consequence of mismatches occurred in CATV system cause ghost image. Their disturbing effect depends not only on the magnitude of reflection coefficient but on time delay of reflected signal too.

Radiated interference field strength generated by *the emission* of RF signals used in distribution network is a very important parameter of CATV system. Each building element of the system has to shield in a measure that the power radiated by

[29] CTB - **c**omposite **t**riple **b**eat: composite third degree distortion is by definition the resultant of third degree discrete distortion products fell into a narrow, 1 kHz band of one TV channel (in case of AM-VSB analogue TV channels its disturbing effects is the most significant if this band is near to picture and colour subcarrier).

[30] CSO - **c**omposite **s**econd **o**rder: composite second degree distortion is by definition the resultant of second degree discrete distortion products fell into a narrow, 1 kHz band of one TV channel.

the system do not exceed the limit values prescribed in MSZ EN 50083-2 standard[31] in case of RF signal levels occurred inside the network [4.8.6]. For example this limit value is 20 dBpW for the frequency band from 30 MHz up to 1000 MHz. The effective shielding also gives immunity for the system against outer RF fields.

The isolation between any two subscriber's outlets protects the receivers against mutual disturbances.

Limit values related to system performance can be found in MSZ EN 50083-7 standard[32] [4.8.7].

Data transmission signals of CATV system must not deteriorate the quality of program service and vice versa. Two important parameters of digital signal transmission are *the bit error rate* (BER) and *the jitter*. The prescribed value of BER in CATV networks depends on service type. For interactive program services its values $1 \cdot 10^{-10}$ [4.8.8] and $1 \cdot 10^{-4}$ [4.8.9]; for data transmission services $1 \cdot 10^{-8}$ in forward- and $1 \cdot 10^{-6}$ in return path [4.8.10], [4.8.11]. The needed carrier-to-noise ratio to the realization of required BER value depends on the chose modulation. In cable modems used for data transmission the manufacturers apply generally 64QAM modulation in forward path and QPSK modulation in return path. An exception is the cable modem manufactured by Terayon firm, which uses modulation of spread spectrum. The QAM modulation has excellent spectral effectiveness (the required bandwidth for transmission of one channel is narrow compared to other modulations); the choice of QPSK modulation is a good compromise between the spectral effectiveness and the reliability of data transmission through noisy channel [4.8.12]. Typical *effective data transfer rates* are 1.92 Mbit/s for a return channel using QPSK modulation with 2 MHz bandwidth, 16 dB carrier-to-noise ratio (BER=$1 \cdot 10^{-9}$); and 23 Mbit/s for forward channel using 64QAM modulation with 6 MHz bandwidth, 23 dB carrier-to-noise ratio (BER=$1 \cdot 10^{-9}$).

[31] MSZ EN 50083-2 standard is the Hungarian equivalent of CENELEC EN 50083-2 standard.

[32] MSZ EN 50083-7 standard is the Hungarian equivalent of CENELEC EN 50083-7 standard.

4.8.4. CATV Data Transmission

Services realized on CATV computer network expect the fulfilment of different requirements from the network. The most important technical requirements are related to the service of different type traffic, the correspondence to several network protocols, the effective and rightful access to commonly used medium, as well as the economical and effective IP[33] address-handling [4.8.13]. This chapter deals only with the technical questions of physical layer for CATV data transmission.

CATV data transmission system is shown in Figure 4.8.2. Its parts are: *two data sources and data sinks*, *cable modems* and *the two-way CATV network*. In Figure 4.8.2 personal computers (PC-s) are as data sources and data sinks; cable modems operate as matching units between the two-way CATV network and PC-s.

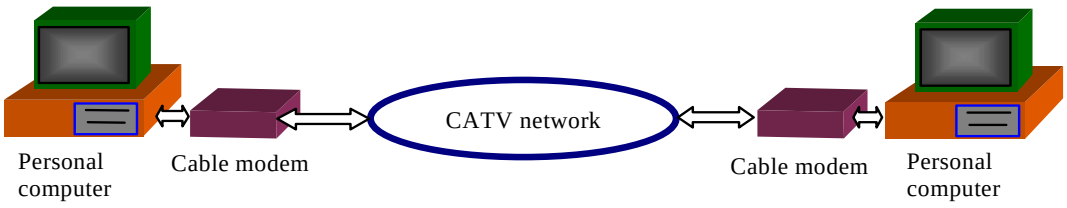


Figure 4.8.2 CATV data transmission system

The frequency band used for data transmission in return path (direction of data stream from subscriber to head-end) is that part of entire return frequency range with 25, 45 55, 60 MHz bandwidth (in case of lowsplit, its value depends on splitter frequency) in which the data transmission channels are placed. In case of FDM multiplexing the frequency band used for data transmission in return path depends on the bandwidth of one data transmission channel, the number of subscribers wanted to connect into this service and the way of media-access. System performance for return path is described in MSZ EN 50083-10 standard[34] [4.8.14]. The planning of return path for hybrid fibre-coaxial CATV networks is discussed in detail by [4.8.15] reference.

In case of symmetrical traffic the bandwidth of data transmission in forward path (direction of data stream is from head-end to subscriber) must be equal to the

[33] IP - **I**nternet **p**rotocol

[34] MSZ EN 50083-10 standard is the Hungarian equivalent of CENELEC EN 50083-10 standard.

bandwidth of data transmission in return path. To put the frequency band for data transmission above 630 MHz (band of digital program distribution) is a logical thing in case of CATV networks with mixed (analogue and digital) program distribution. This can be done because most of cable modem is able to receive forward path signals arrived in the frequency band from 60 MHz up to 862 MHz.

Cable modem system for Internet access will be shown as an example for data transmission service (see Figure 4.8.3). This cable modem system consists of three main parts beyond the devices needed to the connection to Internet. These are: *cable modem* at subscriber, *central unit* at CATV head-end as well as *software for network and subscriber management*.

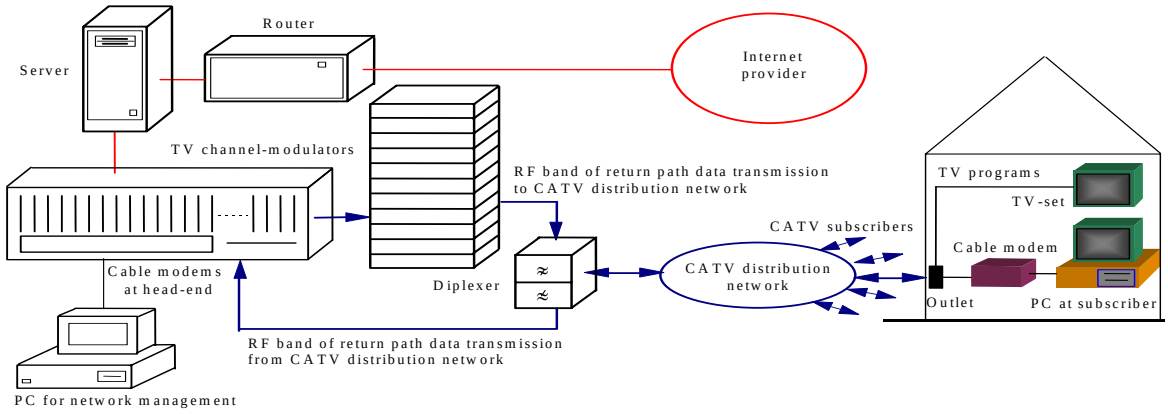


Figure 4.8.3. Cable modem system for Internet access

4.8.5. The Future of Broadband Cable Networks

Nowadays the modern CATV systems regarding distributed programs use analogue-digital technique and HFC technology in their distribution network, consequently they may be considered mixed networks from both points of view. This technique and technology would permit of several services on CATV networks used as broadband transmission media from telephone and NVOD to videoconference, etc. Their advance is not restrained by technical possibilities but considerations of economic efficiency, namely it is determined by *the solvency of people who want these services*. Applying only digital technique in CATV systems and fibre to the subscribers' home in their distribution networks is not technical but economic question. With development of circuit technology and spread of devices made by this new technology the cost of equipments built up these devices will decrease so much,

that services realized by present expensive technical solutions will be accessible for numerous subscribers in the future.

Increase in subscribers' number, integration of small systems into large networks, widening of services involve very quick development of *subscriber- and network-management*. Such kind of service gets also into the sphere of services, which necessitates the reliability and quality (QoS[35]) usual in professional telecommunication. These requirements can be only satisfied by application of *new network topologies*, *standby-equipments*, *parallel transmission paths*, *error correction coding* and *modulations of noise tolerated*. Regarding modulations the use of digital technique has also great importance (see 4.8.1 chapter).

The widening of services raises the question of security regarding to both data transmission and program distribution. The program owners apply encryption in order to avoid program stealing. *Standardization of encryption procedure* and *introduction of uniform technical solution in cable modems* of CATV data transmission systems will decrease the price of equipments needed for realization of services.

Demand on services is not only determined by solvency of people wanted these services but their cultural level too, that is to say that social development influences on the advance of broadband cable systems indirectly.

References

[4.8.1] Géher, Károly: Telecommunications, Chapter 21.5, pp. 252-258, In Hungarian: Híradástechnika, 21.5. alfejezet, Műszaki könyvkiadó, Budapest 2000., ISBN 963 16 3065 X, 252-258. oldal

[4.8.2] Digital Video Broadcasting (DVB), DVB Homepage, (DVB and list of connected standards) http://www.etsi.org/broadcast/dvb.htm , April 29, 2001.

[4.8.3] HDTV FAQ - January 31, 2001,

http://www.nettable.com/HDTVFAQ.HTM , 3 May 2001.

[4.8.4] Advanced Television Standards Committee, ATSC Approved Standards, 11 January 2001, http://www.atsc.org/Standards/stan_rps.html , 3 May 2001.

[4.8.5] Ciciora, Walter; James Farmer; David Large: Modern Cable Television Technology: Video, Voice, and Data Communications, 1st edition, 912 pages, Morgan Kaufmann Publishers, San Francisco, California, 1998 December, ISBN 1 55860 416 2,

[4.8.6] EN 50083-2:1995 Cabled distribution systems for television and sound signals, Part 2: Electromagnetic compatibility for equipment (CENELEC standard), Hungarian equivalent:

[35] QoS: **Q**uality **o**f **S**ervice

MSZ EN 50083-2:1997, Kábeles kép- és hangjelelosztó rendszerek, 2. rész: Berendezések elektromágneses összeférhetősége, MSZT 1997. október, 36 oldal
[4.8. 7] EN 50083-7:1996 Cabled distribution systems for television and sound signals, Part 7: System performance (CENELEC standard), Hungarian equivalent: MSZ EN 50083-7:1998, Kábeles kép- és hangjelelosztó rendszerek, 7. rész: Rendszerjellemzők, MSZT 1998. január, 62 oldal
[4.8.8] IEC 100D/46/NP Cabled distribution systems for television and sound signals, IEC Standard proposal, October 1997.
[4.8.9] ETS 300 800, Digital Video Broadcasting (DVB) Interaction channel for Cable-TV distribution system, ed1, European Telecommunication Standard, June 1998.
[4.8.10] IEEE 802.14/a. Cable-TV access method and physical specification, Draft 2, July 1997.
[4.8.11] MCNS Data-Over-Cable Service Interface Specification, September 1997.
[4.8.12] Mátay, G.: Data Transmission on Cable Television Networks, 2nd ELMAR International Workshop on Video Processing and Multimedia Communications, June 28-30, 2000, Zadar, Croatia, Proceedings VIPromCom-2000, pp. 59-63.
[4.8.13] Gróf, Róbert: Cable Modems in the interactive CATV network, Diploma work, Budapest University of Technology and Economics, Department of Microwave Telecommunication, 1997, In Hungarian: Kábelmodemek az interaktív kábeltelevízió hálózatban., Diplomaterv, BME-Mikrohullámú Híradástechnika Tanszék, 1997.
[4.8.14] EN 50083-10:1999 Cable networks for television signals, sound signals and interactive services. Part 10: System performance for return path, Hungarian equivalent: MSZ EN 50083-10:2000, Televíziójelek, hangjelek és interaktív szolgáltatások kábeles elosztóhálózatai. 10. rész: A visszirány rendszerjellemzői, 30 oldal
[4.8.15] Raskin, Donald; Dean Stoneback: Broadband Return Systems for Hybrid Fiber/Coax Cable TV Networks, Prentice Hall, New Jersey 1998, 297 pages, ISBN 0 13 636515-9.

4.9. Terrestrial mobile networks

Andrea Schmitterer-Bausz, dr. (4.9.0.-3.), Károly Fiala, dr. (4.9.3.), Gyula Simon, dr. (4.9.4-8.), authors

István Maradi, reviewer

Mobile communications is a general term to cover different mobile systems developed for various user needs and consequently providing different services. The common feature of mobile systems is to offer mobility, i.e. the freedom to move and stay in touch, though the extent of mobility depends on the system type. Mobile terminals have access to the network through the air interface, and the system performs mobility management including location registration and up-date, and handover.

Different types of terrestrial mobile systems are listed in Table 4.9.1, and we cite some widely known standards as examples. Mobile satellite systems, providing wide area or even global coverage are described in Subchapter 4.10.

Technical evolution resulted in the development of several generations within mobile system categories. The first mobile system had been launched in St. Louis (USA) in 1946, however wider spread was experienced only during the eighties by the emergence of the first generation (1G) systems through the evolved integrated electronic devices and microprocessors. 1G networks are called analogue systems because of the use of analogue radio technology, however the applied switch and transmission technologies are digital. Second generation (2G) systems emerging in the nineties use already digital radio technology.

1G systems were primarily based on proprietary developments, though some of them became de-facto standards, such as the NMT system developed in the frame of a regional North-European co-operation, introducing international roaming first.

The planning process of 2G was already led by the aims of the European Community for market liberalisation, strengthening competition by creating unified European standards. The rest of the world however did not harmonised 2G

standardisation, so there might be invariably several mobile standards within each system category simultaneously.

In this chapter we present the most important terrestrial mobile systems. Public mobile radiotelephone systems are introduced through the European GSM standard and its further developments towards 2,5G and 3G, and trunked radio systems and paging systems are described.

Terrestrial mobile systems		
System category	*1G*	*2G*
Public land mobile networks (PLMN), also known as cellular networks	*NMT450/900, TACS, AMPS*	*GSM, PDC, IS-95 CDMA, IS-136 TDMA*
Trunked radio networks, also known as private or public access mobile radio (PMR/PAMR)	*MPT1327*	*TETRA, TETRAPOL*
Private mobile data networks	*MOBITEX, MOBICOM*	*-*
Paging networks public systems private, on-site systems	*POCSAG*	*ERMES*
Public terrestrial flight telecommunications system	*-*	*TFTS*
Cordless telephony (systems)	*CT0, CT1*	*CT2, DECT*
Wireless LAN (WLAN)	*-*	*HIPERLAN1, 2 IEEE 802.11*
Short range devices		*DSRR Bluetooth*

Table 4.9.1 Overview of terrestrial mobile systems with some 1G and 2G examples

4.9.1. Public land mobile networks - the GSM system

GSM (Global System for Mobile Communications) is a Pan-European digital cellular system standardised by ETSI. During the elaboration of GSM the target was European harmonisation. Hereby all mobile stations can be used in all participating countries (international roaming), and the equipment from different vendors are compatible and able to interwork with each other, so the equipment prices and consequently the service tariffs come to a significantly lower level than ever. Therefore network functions and interfaces are specified by the GSM standards, however hardware requirement is not defined.

In Europe the primary spectrum allocated for GSM is in the 900 MHz band (890-914/935-959 MHz). Later a GSM version operating in the 1800 MHz band (1710-1785/1805-1880 MHz) was developed as well under the original name of

DCS1800 (Digital Cellular System). It is consistent with the GSM900 standards apart from some differences in the radio part due to the different band applied. The GSM version operating in the 1900 MHz band was prepared for the USA in 1995.

The elaboration of GSM standards is a continuous work. The publication of the standards was organised in two phases because of the complexity of the system in 1991 and 1995, since then Phase 2+ with new capabilities was released annually until R99 (R, Release). New developments in data transmission establish already the basis of future mobile systems building on GSM.

As a result of the successful European concept GSM is spread all over the world. At the end of April 2001 more than 400 GSM networks were in operation in 169 countries, and the number of subscribers exceeded half billion. Thus GSM holds 70 % share of the digital cellular market worldwide.

GSM architecture

Figure 4.9.1 shows the architecture of the GSM system, which consists of four subsystems [4.9.1]:

Mobile Station (MS)

Subscribers use mobile stations to access the services through the radio interface. Each subscriber has a Subscriber Identity Module (SIM), which is a removable chip-card and forms part of the MS. The SIM stores subscription data and fulfils security tasks as well.

A mobile station can act as a Terminal Adapter (TA) as well, providing connection towards other terminals such as fax equipment or personal computers, depending on the services supported by the MS.

Base Station Subsystem (BSS)

The Base Station Subsystem provides continuous radio connection for the mobile station to access the network. Functional entities of the BSS are the following:

- Base Transceiver Station (BTS)

The base station is basically a radio transciever with antennas, which provides coverage within a given area, namely in the cell.

- Base Station Controller (BSC)

The Base Station Controller is a smaller capacity switch, which controls all radio related tasks: the channel assignment for base stations and mobile stations, and the handover process between cells. Since the standardisation of the A-bis interface between BSC and BTS is not fully detailed, the BSS can not be broken up.

However BSS and NSS are connected by the standardised A interface, so the radio and network subsystems of different vendors interwork with each other.

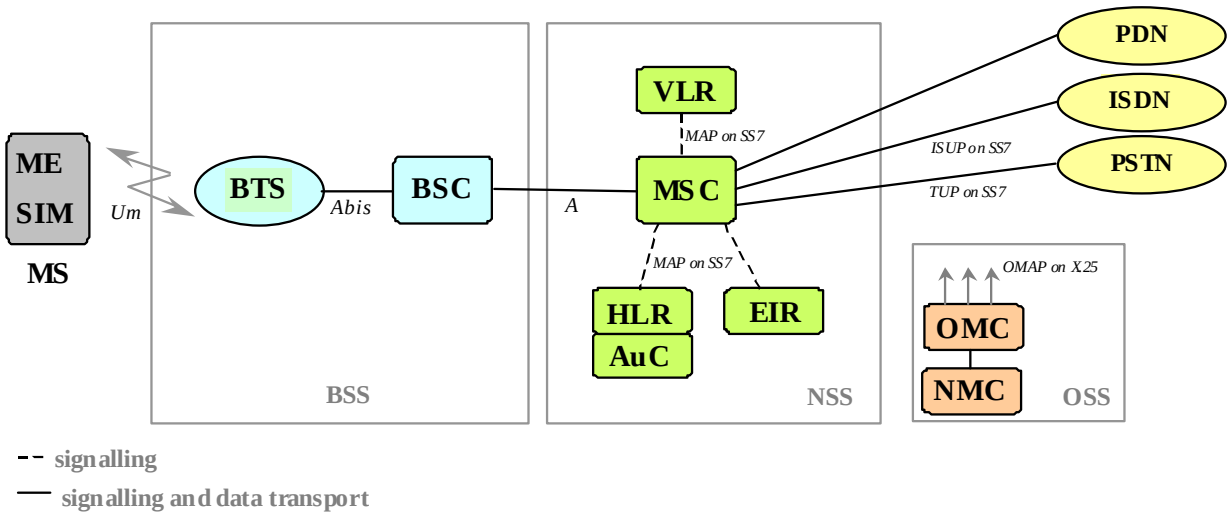


Figure 4.9.1 GSM system architecture [4.9.2]

Network and Switching Subsystem (NSS)

The NSS performs switching functions of GSM, controls call set up, provides interface to other networks and consists of databases for subscriber identification and tracking, and for terminal check. Information exchange between the units of the NSS is based on SS7. Elements of the NSS:

- Mobile services Switching Centre (MSC)

The base of the MSC is an ISDN switch, complemented with additional functions to manage radio resources and subscriber mobility (i.e. location registration and handover). The GSM network interconnects with other telecommunications networks through the Gateway MSCs (GMSC), which route the incoming calls towards the serving MSC. The number of GMSCs in a network is based on the operator's decision, even each MSC can be a GMSC as well. The Interworking

Function (IWF) is part of the GSMC, and performs protocol conversion between GSM and fixed networks.

- Visitor Location Register (VLR)

A VLR database belongs to each MSC, and it keeps records of those subscribers, who are actually staying in the service area of the given MSC.

- Home Location Register (HLR)

The mobile operator keeps records of each of its subscribers in the HLR database. These records contain subscription data, subscriber identities, service restrictions and the routing information necessary to reach the subscriber at the actual place of staying.

- AUthentication Centre (AuC)

The AuC database records an identity key to each mobile subscriber registered in the HLR. It is used for subscriber authorisation and for the encryption of the radio channel, in order to prevent unauthorised use.

- Equipment Identity Register (EIR)

The EIR database records the International Mobile Equipment Identity (IMEI) of mobile stations. There are three different lists in the EIR. The white list contains mobile stations with type approval, the grey list records mobile stations under observation, and the black one lists blocked mobile stations due to theft or problems in operation.

Operation & maintenance SubSystem (OSS)

The Operation & maintenance SubSystem has a hierarchical structure. The elements of the GSM network have an X.25 data connection to the regional Operation & Maintenance Centre (OMC). The OMC provides information on the network status, and provides possibility of tuning several system parameters. The Network Management Centre (NMC) provides central supervision and network management functions. In a GSM network a single NMC is needed to control the local OMCs. The staff of the NMC is in charge of issues effecting the whole network and requiring national co-ordination, and deal with traffic management, while the regional OMC groups handle short term local business.

4.9.2. Data developments in the GSM network – 2,5G network architecture

ISDN had a significant impact on the elaboration of GSM services and signalling protocols. GSM was defined as a “voice+data” system already by the basic system concept, and the aim was to ensure maximal flexibility to allow ISDN-type services. Standardisation of GSM data services was going on from the very beginning. Mobile originated Short Message Service (SMS) and fax transmission at the rate of 9,6 kbit/s were available since 1994, and circuit switched data service also at the rate of 9,6 kbit/s was ready in 1995. The next step was a new coding scheme providing 14,4 kbit/s transmission rate per time slot within the frame of Phase 2+. Then **HSCSD** (High Speed Circuit Switched Data) came in 1998, which provides maximum 57,6 kbit/s transmission rate by bundling four time slots, and it supports asymmetric traffic. For the introduction of HSCSD only a software upgrade is necessary on the network side.

In the meantime work on 3G has been started in parallel with the GSM developments, and the service requirements defined there reacted on the standardisation of GSM Phase 2+. Since GSM with the new, enhanced capabilities will ensure some of the 3G expectations to a certain extent, and the new network elements may be utilized for the introduction of 3G, the new GSM data solutions are called 2,5 generation as well.

GPRS (General Packet Radio Service)

GPRS is a new bearer service created by the expansion of the existing GSM architecture and by the introduction of new channel coding schemes. GPRS provides packet switched data services. GPRS means the first step towards the integration of the mobile and IP based services, since several IP standards (IETF RFC) were used to form GPRS.

The maximum bit rate available in GPRS is 171,2 kbit/s in principle, by bundling eight time slots and using the channel coding scheme providing the lowest data protection. However in practice 40,2 kbit/s is available on the radio channel for the time being, using CS2 coding and bundling 3 time slots. This means 30-33 kbit/s on the application layer level, depending on the application. Through packet

switching GPRS provides a continuous virtual connection (always-on), and charging according to connection time as used in circuit switched systems is replaced by tariffs based on the volume of sent packets.

GPRS traffic is separated from the traditional circuit switched voice and data traffic. Figure 9-4-2 shows GPRS architecture and its relation to the GSM core network. In order to bypass the existing GSM core infrastructure new nodes are necessary based on IP routing equipment. HLR is enhanced with the GPRS register, so it contains the GPRS subscription data as well. The new fixed network entities:

- the Servicing GPRS Support Node (**SGSN**) is connected to the BSS through the Gb interface based on Frame Relay, and performs mobility management, security functions (authorization, encryption) and packet routing,
- the Gateway GPRS Support Node (**GGSN**) provides connection towards other networks, performs security and billing functions, and dynamic IP address assignment. GPRS Phase 1 supports standardised TCP/IP and X.25 protocols. The Gn interface between the GGSN and SGSN is based on IP.

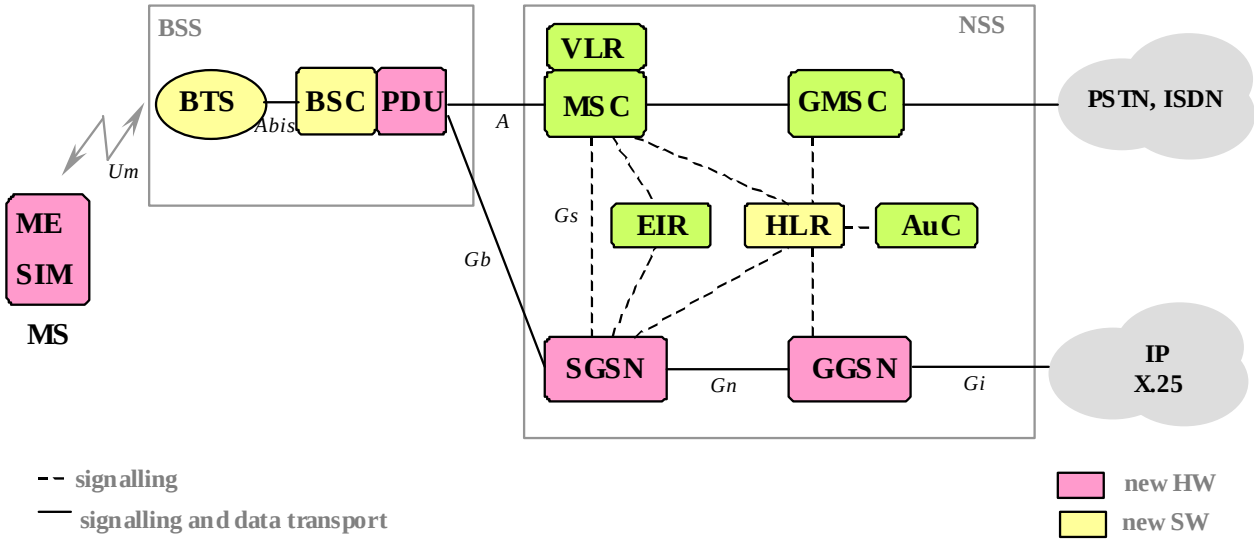


Figure 4.9.2 GPRS architecture [4.9.3]

In the base station subsystem:

- the BSC HW is enhanced with a Packet Control Unit (PCU), which is connected to the SGSN through a new interface, and a software upgrade is necessary also,
- the BTS ensures the dynamic resource allocation between GSM and GPRS with a new SW.

To access the service new mobile stations are necessary with GPRS capability. The equipment is ranked into three classes: Class A and B terminals provide GSM/GPRS capability, but only Class A is able to handle circuit and packet switched traffic simultaneously. Class C has pure GPRS capability only.

EDGE (Enhanced Data rates for GSM Evolution)

To be complete we have to mention EDGE as well, which provides significantly higher data rates on the radio interface for slowly or non-moving subscribers by introducing new modulation and coding methods. Consequently EDGE means installation of new radio units, i.e. new base stations or at least new TRXs to the GSM radio network.

In the frame of EDGE the enhanced circuit switched data service (ECSD) will offer 28,8 kbit/s rate per time slot, the limitation of 64 kbit/s is derived from the A interface. The packet switched EGPRS will provide maximum 384 kbit/s rate in case the speed of the moving subscriber is below 100 km/h. EDGE forms an integral part of the GSM evolution, and it might be an advantageous solution for rural areas, where the provision of UMTS coverage is uneconomical.

4.9.3. 3 G mobile networks

The elaboration of third generation mobile systems providing a variety of voice, data, and multimedia services is co-ordinated by the ITU in the frame of the IMT-2000 family of standards since 1986. UMTS (Universal Mobile Telecommunications Systems) is a member of this family, and its standardisation was started by ETSI in 1991, then carried on by 3GPP (3G Partnership Project) since 1999. 3GPP aims to create 3G standards based on the evolved GSM core network and on UTRA (UMTS Terrestrial Radio Access) specified by ETSI.

UMTS architecture

While forming the UMTS concept it became an important point to take the existing GSM/GPRS networks into account. It was necessary to consider GSM precedents both from political and economical point of view, noting the subscriber base acquired, the huge investments in development and installations, and the long pay back period in the competitive environment established.

The keys to 3G success are the utilization of the already existing infrastructure and developments based on evolution, which require more reasonable investments.

The UMTS core network (CN) is built on GSM and GPRS network entities, ensuring by this the use of former equipment through pre-defined developments. UMTS core network is standardised in several steps.

Release 99 describes a hybrid architecture with separate circuit switched (CS) and packet switched (PS) domains in the core network. These domains are based on the GSM MSC and GPRS nodes extended with UMTS specific developments (E-MSC and E-GSN) as it is shown in Figure 4.9.3. In the UMTS radio access network the base station is called Node B, and the name of the base station controller is Radio Network Controller (RNC).

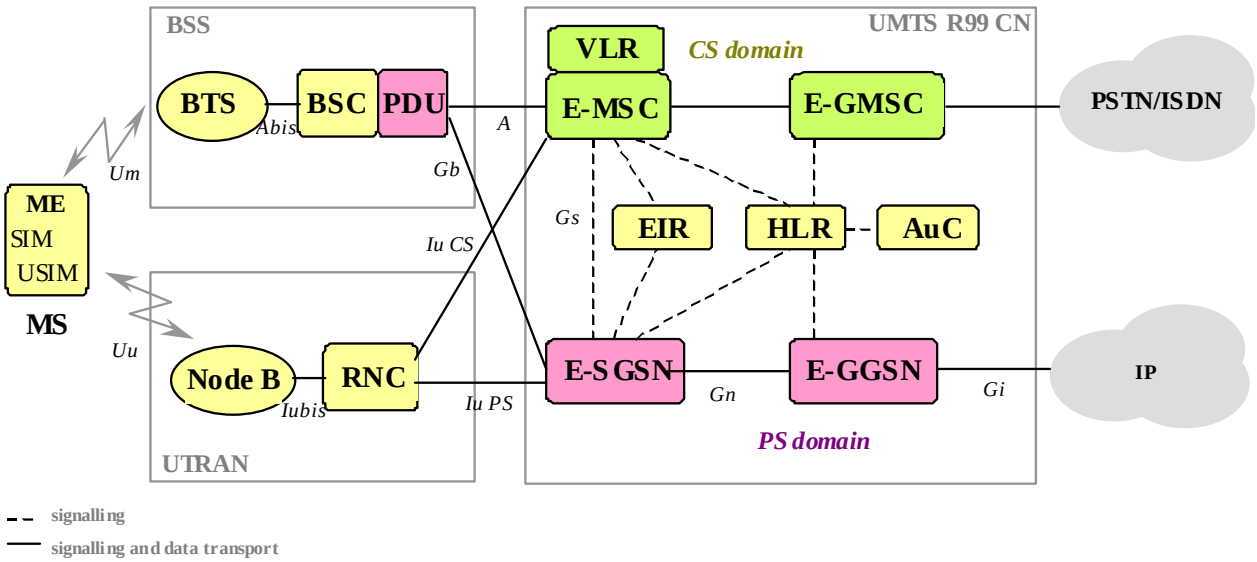


***Figure** 4.9.3 UMTS R99 architecture [4.9.4]*

The solution under preparation within the frame of Release 5 consists of a distributed architecture based on IP (referred as “All IP based core network”), and its standardisation is scheduled to be ready by December 2001 [4.9.5]. The main points are to divert legacy GSM/UMTS voice traffic to the packet switched domain, consolidate the core network towards a general purpose multi-service connectivity network, separate the transport and call control functions, and to introduce a wide range of real-time multimedia and information services applying IETF SIP (Session

Initiation Protocol), which requires a new, fully integrated IP-Multimedia (IM) subsystem in the core network.

Entities of the new IM subsystem:

- the Call State Control Function (CSCF) is a gateway for incoming calls, it performs call control and address handling,
- the Media Gateway Control Function (MGCF) is a PSTN/PLMN termination point, which performs protocol conversion between ISUP and the call control protocols of the IM subsystem,
- the Media Gateway Function (MGW) is a PSTN/PLMN transport termination point responsible for resource management,
- the Multimedia Resource Function (MRF) provides multi-party and multimedia conference functions and handles the necessary bearer control tasks,
- the Transport Signalling Gateway Function (T-SWG) is a PSTN/PLMN termination point, which maps call related signalling from/to PSTN/PLMN on an IP bearer,
- the Roaming Signalling Gateway Function (R-SGW) supports roaming between 2G/R99 and 3G/R5 CS and PS domains.

Further new entities:

- the Home Subscriber Server (HSS) provides HLR functions according to 3G expectations, with IP based interfaces,
- the MSC server performs the call and mobility control functions of a GSM/UMTS R99 MSC.

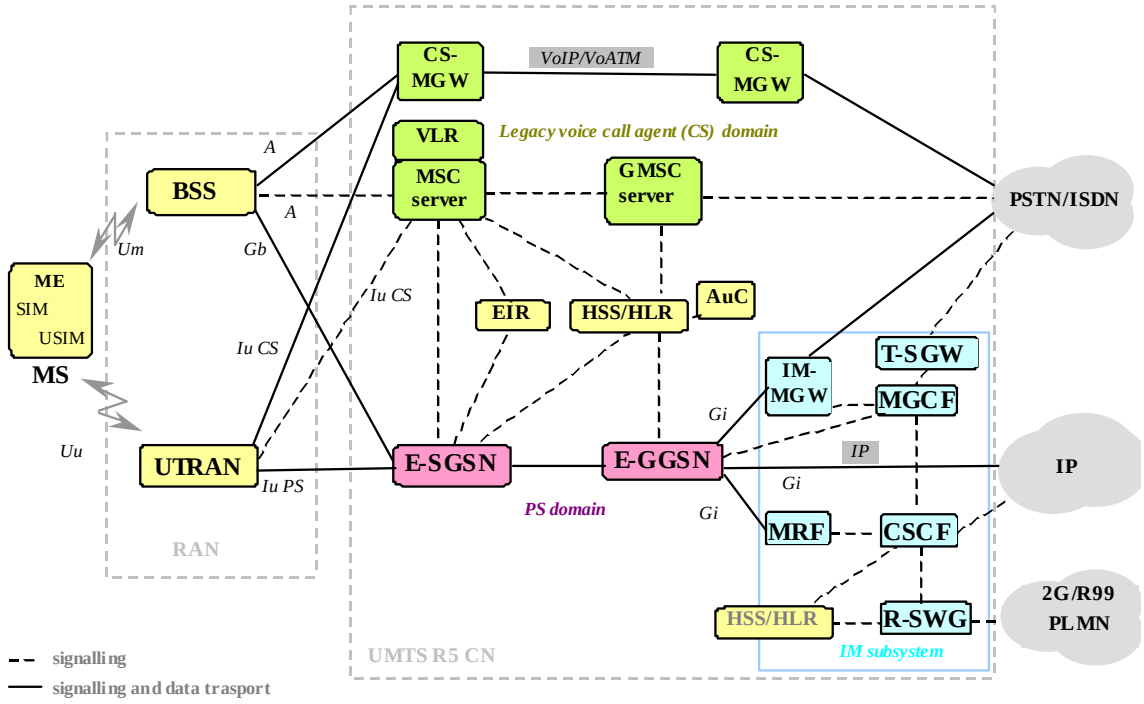


Figure 4.9.4 UMTS R5 architecture [4.9.6]

Provision of the UMTS/IMT 2000 frequency band *(K. Fiala)*

1992 WARC Malaga-Torremolinos: designating the „core" band (1885-2025 and 2110-2200 MHz)

In 1992 in Malaga-Torremolinos the World Administrative Radio Conference (WARC) defined the frequency bands (S5.388) of the third generation system (FPLMTS) within the 2 GHz range by designating altogether 230 MHz „core" bands, of which 60 MHz belongs to the satellite component and 155 MHz will be the frequency band of the terrestrial 3G systems (15 MHz is in the DECT band). The core band is *1885-2025 and 2110-2200 MHz,* the band used by terrestrial systems is *1900-1980 MHz, 2010-2025 MHz and 2110-2170 MHz.* On June 30, 1997 the European Radiocommunications Committee passed a resolution (ERC/DEC/(97)07) according to which by January 1, 2002 at least 2x40 MHz must be provided for purposes of UMTS, when it is substantiated by market demand. The UMTS Forum played an important role in the fact that DG XIII. of the European Committee instructed CEPT to revise this resolution and stipulated that the full 155 MHz band must be made available. The UMTS Forum conducted comprehensive studies to review the frequency demands (Reports No. 5., 6. and 7.) substantiating that the 2x40 MHz wide frequency band is not sufficient for even the launch planned in 2002. The currently effective ERC Resolution – passed on March 28, 2000 (ERC/DEC/(00)01) – obliges Administrations commiting themselves to this Resolution to make to whole 155 MHz terrestrial UMTS frequency range (1900-1980 MHz, 2010-2025 MHz and 2110-2170 MHz) available in the CEPT countries at the latest by January 1, 2002, for purposes of terrestrial UMTS and other systems being part of the IMT-2000 standard family. (The Hungarian Administration did not commit to implement Resolution ERC/DEC/(00)01). The UMTS Forum recommends the designation – in the framework of comperative bidding – of 2x15 MHz paired and 5 MHz unpaired frequency blocks to Administrations as the preferred minimum frequency requirement per public UMTS operator in the initial phase. Resolution ERC/DEC/(99)25 – passed on November 29, 1999 – also includes the frequency plan, in relation to the core band. Based on the National Footnote of H 141 of the Hungarian National Table of Frequency Allocations, the 1900-1980 MHz, 2010-2025 MHz and 2110-2170 MHz bands are planned for fixed and mobile UMTS terrestrial

applications. According to National Footnote of H143 bands 1980-2010 MHz and 2170-2200 MHz are planned for the satellite applications of UMTS.

2000 WRC Istanbul: additional frequency bands (2500-2690 MHz, 1710-1885 MHz, 806-960 MHz)

At the Istanbul World Radiocommunication Conference – 2000, resolution was passed on additional frequency bands for IMT-2000 systems. The UMTS Forum played a major part in increasing frequency possibilities for UMTS as the Spectrum Aspect Group (SAG) of UMTS Forum forecast 403 MHz for 2005 and 582 MHz by 2010 for purposes of terrestrial UMTS at the most densely populated areas. In its Report No. 7 the UMTS Forum – in addition to the second generation systems' bands (altogether 240 MHz) – recommended to designate the additional 187 MHz band. The additional bands enable UMTS to emerge in multiple bands. Lower frequency bands in theory offer the possibility of national coverage and the introduction of UMTS in rural areas, within developing countries. The preferred additional frequency band by CEPT is 2500 – 2690 MHz. The European Committee mandated CEPT to harmonise the additional frequency bands in Europe.

4.9.4. Terrestial trunked radio networks

These non-public networks of typically cellular structure consist of two distinct groups. The private networks (Private Mobile Radio, PMR) are in the first group, the examples of users include the emergency or governmental organizations. The network operator of the second group however provides virtual private network services (Public Access Mobile Radio, PAMR) for user (subscriber) groups (e.g. transport companies, public utilities, etc.). In both of the above cases the common use of the radio resources (i.e. trunking) results in advantageus features as far as the traffic capacity is concerned. The possibility of half-duplex connections is typical. Generally fix stations (dispatcher stations) are integral parts of the system. Group (point-to-multipoint) calls are very often used, where the transmitting party can be changed, and the actual transmitted information are intended to be delivered to each member of the group. The role of the sending party can be varied in time.

Important requirements for emergency organizations are the extremily short (in the range of some hundred milliseconds) connection setup time, the proper ciphering and the possibilty of emergency calls.

Two of the second generation digital systems will be presented below.

The TETRA (TErrestial Trunked RAdio) terrestial trunked radio system

In Europe there are frequency bands of 10 MHz duplex spacing (380-390 and 390-400 MHz, respectively for PMR, 410-420 and 420-430 MHz, respectively for PAMR purposes) for the standardized TETRA system introduced by the European Telecommunications Standards Institute for both PMR and PAMR applications.

Within this frame there are three system standards: Voice plus Data (V+D), Packet Data Optimized (PDO) without the possibility of voice communication, and the Direct Mode of Operation (DMO).

The equipment according to the V+D specifications are providing a great choice of such bearer services, teleservices and supplementary services, which meet the requirements set by both the voice and the data transmission. The advantages resulting from the flexibility of the V+D system are more attractive for several applications than the slightly higher speed of PDO stemming from being optimized for data. As far as the physical radio air interface layer is concerned (e.g. modulation, carrier frequencies, etc.) the specifications for V+D and PDO are the very same, however at this level there isn't any possibility for interoperability. The DMO is providing a direct mobile-mobile connection (even without any base station), if the mobile concerned is outside the coverage area of the system, or if there is a requirement for a more secure connection within the coverage area. A given mobile operating in a TETRA V+D network is also able to get connected to a mobile functioning in DMO The mobile equipment therefore should watch both systems (dual watch, see Figure 4.9.5)

At the design of TETRA V+D a feature of prime importance was represented by the suitability for multimedia applications, thus. transmission and reception of voice, data, images, etc is possible for all (suitably designed) mobile stations in the system. The main limit is the transmission speed.

PMR V+D systems (combined with DMO) are to be deployed first.

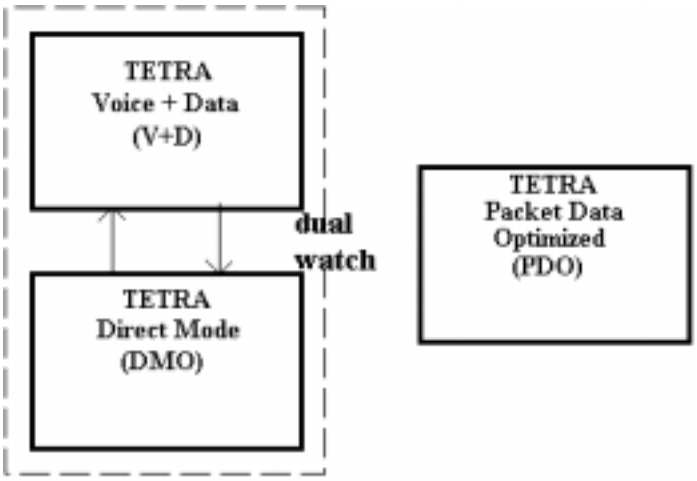


Figure 4.9.5: Modes of operation for TETRA

The digital standardized system TETRA includes mobile stations (the vehicular set is called mobile, while the portable one is called handportable), wired stations and the central infrastructure (consisting e.g. of base stations and the mobile switching center as essential parts). Some interfaces within the system are standardized (the ones in the infrastructure aren't!). The architecture of the system together with the standard interfaces is shown in Figure 4.9.6.

The type of the modulation is π/4DQPSK, there are four possible levels within a symbol interval, a symbol therefore is representing two bits. Each dibit is represented by a phase change depending on the dibit value referred to the phase of the last transmitted symbol. The access type is a time division type (TDMA), a frame consists of four timeslots. Thus one carrier frequency yields four transmission channels. The frequency spacing is 25 kHz. The net bit rate of a channel is 7,2 kbit/s, by bundling up four timeslots a transmission speed of 28.8 kbit/s would be reached. Properly deployed systems are providing the possibility for handover and roaming, too.

TETRAPOL

A PMR system can also be deployed on the basis of TETRAPOL, which is a system initiated by companies. The access is of frequency multiplex type (FDMA), the channel spacing is 12,5 kHz, the modulation is GMSK with a speed of 8 kbit/s.

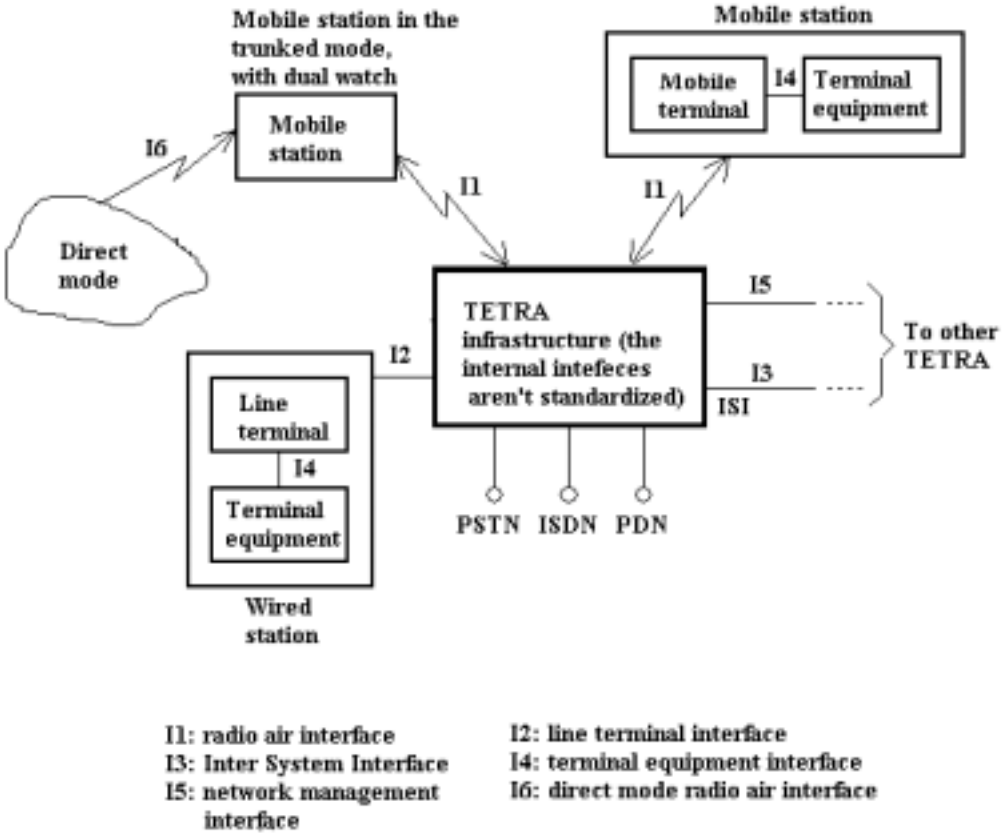


Figure 4.9.6. *The architecture of the system and the standard interfaces*
(Note: the interface of lawful interception is also standardized)

Mthe main advantages include an end-to-end efficient ciphering scheme, the disadvantage is the lack of handover.

4.9.5. Paging networks

The radio paging systems are providing unidirectional information transmission. The messages are transferred to the subscriber's reciever of small power consumption (battery powered) and of small size as a transmission by one of the radio transmitter of the cellular system, and the information is appearing on the display of the receiver. The sending party may initiate the related message at the service center (mainly by phone) or directly from a computer. The standardized Trans-European system is called European Radio MEssaging System (ERMES). The widespread use of public mobile radio systems and specifically the service called SMS within these led to a considerable decrease in the subscription number.

4.9.6. Cordless systems - DECT

The DECT (Digital Enhanced Cordless System) is a FDMA/TDMA/TDD based system- The frequency spacing is 1.728 MHz, the bit rate is 1.152 Mbit/s. The frame consists of 12 uplink and 12 downlink timeslots. The voice coding is ADPCM of 32 kbit/s speed. In addition to the cordless access to the subsribed wired telephone network (in the range of 200m radius) DECT is also frequently used in wireless local loops (connecting the local exchange and the subscriber) as well as in Wireless Private Automatic Branch eXchanges (WPABX). The compability in any special application is granted by more specific, so called profile standards.

4.9.7. Wireless LAN - 802.11

The operating freqency range of the wireless LAN falls into an Industrial, Scientific and Medical (ISM) band (e.g. 2400-2483,5 MHz), e.g. the mentioned frequency range is used as well by the microwave heaters. The protection against interferences therefore is provided by spread spectrum techniques (based on either direct sequential or fast frequency hopping type). The maximum trasmitted power is 100 mW resulting in a maximum distance of about100 m.

4.9.8. Small area networks - Bluetooth

The system Bluetooth is not a standard, as established by the cooperation iof companies . It can be considered to be a de-facto standard.

Bluetooth is making use of the same ISM frequency band as the 802.11, only a special version (being capable of power control) may have an output power of 100 mW, in other cases the output power should be of lower value. The effective range is therefore reduced to some ten meters. In a duplex case the duplexing is time division based. The operation is slotted, the nominal slot interval is 625 μs. A packet consists of 1...5 timeslot(s). A voice channel is 64 kbit/s synchronous in both directions, the asynchronous data channel may be asymmetric (723.2 kbit/s max. in the direct while 57,6 kbit/s max. in the reverse direction) or symmetric (2 x 433,9 kbit/s max.). In a so called piconet the maximum number of devices is limited to 7. A special device is the master, the members/devices in a piconet use a common channel defined by a slow

frequency hopping sequence. A given device may be member in more piconets resulting in a so called scatternet structure. Bluetooth is representing an universal solution for short-range connections of all kind (e.g. between a mobile phone and a microphone/earphone, between a computer and the related peripheries, etc.).

References

[4.9.1] Mouly-Pautet: The GSM System for Mobile Communications; Cell&Sys, 1992.

[4.9.2] GSM 03.02 v5.1.0 1996-05: Digital cellular telecommunications system (Phase 2+); Network architecture

[4.9.3] GSM 03.60 v5.2.0 1997-12: Digital cellular telecommunications system (Phase 2+); GPRS Service description, Stage 2

[4.9.4] 3G TS 23.002 version 3.3.0 Release 1999: Digital cellular telecommunications system (Phase 2+) (GSM); UMTS Network architecture

[4.9.5] Huber-Weiler-Brand: UMTS, the Mobile Multimedia Vision for IMT-2000: A Focus on Standardisation; IEEE Communications Magazine; Sept. 2000, pp. 129-136.

[4.9.6] 3GPP TS 23.002 v 5.2.0 2001-04: 3GPP TSG SA, Network Architecture (Release 5)

4.10. Satellite networks

Éva Gödör, author

István Hazay, reviewer

Electronic systems on board of satellites can be used for several purposes such as

- research (radio astronomy, radar astronomy, meteorological remote sensing,
- navigation,
- telecommunications.

For these purposes, several specialized systems have been developed, with terrestrial stations or networks used to process, distribute, store etc. the satellite signals resulting e.g. from meteorological observations [4.10.1]. At present, satellite signals are simultaneously utilized in several devices: for instance, the signals originating from navigational satellites of the Global Positioning System (GPS) can be displayed on mobile phones too, and in satellite communication networks, it is not always possible to distinguish the functions of fixed, mobile and broadcasting networks (FSS – Fixed Satellite Service, MSS – Mobile Satellite Service, BSS – Broadcasting Satellite Service). Note also that within a single transmission path connecting two subscribers, several transmission media such as optical fibres, terrestrial microwave links and satellite transmissions can be alternately applied. This is why the concept of convergence and integration is frequently treated in the literature [4.10.2].

In this Section, primarily satellite mobile systems and satellite information gathering and distribution systems are dealt with. Satellite broadcasting systems are surveyed in Sec. 4.7.

4.10.1. Satellite communication

In classical satellite communication systems (e.g. INTELSAT), the network node is the satellite itself, accessed by high-capacity fixed terrestrial stations by utilizing the multiple access principle (e.g. FDMA, later TDMA, CDMA), thus

establishing the Fixed Satellite Service (FSS). According to this service, the satellite is utilized by the communicating terrestrial stations as a transparent repeater station. The transponder on board of the satellite receives the terrestrial signal, shifts the signal frequency, amplifies the signal and radiates it down to the terrestrial station. This kind of connection has an uplink and a downlink part resulting in a single hop, represented by a folded-back signal path. Subsequently, additional functions emerged by executing terrestrial commands to select, from a set of beams (global, spot, zone, hemispherical beam) a suitable one meeting current traffic requirements.

The geographical extension of the network is given by the footprint of the satellilte coverage. This depends on the satellite orbital plane, the type of orbit, the orbital height and the antenna radiation pattern. Assume now that from the two geographical locations that are the terminals of the satellite link, only one of them is within the satelllite footprint. In this case, two satellites are used for the transmission, i.e. two hops are applied. However, these two hops require substantial propagation time, e.g. in case of a geostationary orbit (GEO), one hop requires 250 ms on the average. This means that the 500 ms propagation time of the two hops in each direction renders the voice transmission and the dialogue of the parties rather cumbersome.

However, there is another solution: e.g. following one-hop transmission, the signal can further be transmitted over a terrestrial path (microwave poin-to-point link, optical fibre or copper cable connection). A perspective possibility is the establishment of a link between satellites (Inter Satellilte Link, ISL). This calls for an on-board processing system according to which the satellite acts no longer as a simple repeater station but has also a processing function.

Following the success of FSS, another satellite system was created, called Mobile Satellite Service (MSS), serving subscribers from whom at least one is on the move. The first commercial organization establislhing MSS was INMARSAT. At first, this was responsible for connecting a maritime ship with the mainland, or a pair of briefcase-sized terminals located in planes or cars, anywhere over the globe. The elements of the INMARSAT system are the following: four geostationary satellites with conical global beams, orbiting above the global oceans, terrestrial signal processing stations, and the mobile terminals. Subject to the terminal type, the transmitted information is voice or data, with bit rates in the 1.2 to 16 kbps range. The

mobile telecommunication over satelllite has thus been realized: anyone, anywhere and anytime can be reached. Earlier, the limited bit rate range was no problem, but later, the increasing informatrion content urged the service providers to offer broadband solutions. For example, the Global Area Network (GAN) program has been started by INMARSAT in 2002. In course of this program, global mobile broadband services are offered providing voice, ISDN and IP services over the THURAYA satellites at 144 kb/s bit rate, and subsequently over the INMARSAT-4 system, at 432 kb/s.

The main development trends of satellite systems are economy, usability according to demand, mobility and broadband capability [4.10.3], [4.10.4], [4.10.5]. In addition to geostationary mobile systems, Low Earth Orbit (LEO) and Medium Earth Orbit (MEO) mobile systems have been developed such as IRIDIUM, GLOBALSTAR, TELEDESIC, SKYBRIDGE, New ICO. A common feature of these systems is the use of several tens of spotbeams per satellite creating a terrestrial cellular system. In view of the orbital characteristics of mobile satellite systems, not the user is moving in the cell area, but the cell itself is moving. This is understandable by considering that compared to a vehicle speed of e.g. 170 km/h, the IRIDIUM satellite speed relative to the earh is 28500 km/h. Considering the cells created by the spotbeams, they are either moving together with the satellite (e.g. IRIDIUM system), or they are made to cover a fixed area on the Earth surface by electronic control, as long as the satellite appears over the horizon. This latter system is realized by the ground-based cells of the TELEDESIC system.

A crucial problem of cellular systems is the handover of calls. There are three basic types of handover used in satellite cellular systems [4.10.5], [4.10.6]:

- handover within the satellite (intraspot and interspot handover),
- handover between satellites (intersat handover, e.g. IRIDIUM, TELEDESIC),
- handover between the satellite system and the terrestrial system (e.g. the ICONET backbone system), or handover between the satellite and the Mobile Services Switching Center (MSSC) of the terrestrial GSM system, or handover between GSM base stations over satellite, substituting the terrestrial point-to-multipoint microwave Wireless Local Loop (WLL).

In addition to mobility, broadband capability is an important parameter, enabling, among others, multimedia and Internet services over satellites. For this purpose, a new frequency band has been allocated: this is the Ka-ban (20/30 GHz).

Further, in order to enhance effective on-board processing, a fast packet-switvched transmission method, similarly to the Asynchronous Transfer Mode (ATM), had to be assigned, as e.g. in the TELEDESIC, ASTROLINK, SPACEWAY, CYBERSTAR systems.

The Ka band is not a simple extension of the Ku band (12/18 GHz), because its propagation parameters in the atmosphere are substantially different. Further, there is a substantial difference in the size of the usable antenna. The architecture of Ka band systems are governed by all these properties.

Here are a few applications of broadband ATM-based satellite systems [4.10.4].

- transmission of low-speed data/picture/short message/paging,
- ineractive computing,
- dissemination of information (e.g. stock exchange data),
- transmission of bulk data (e.g. ftp, http),
- video conferencing, net-conferencing,
- multicasting,
- transmission of large data files, e-mails to collective addresses,
- broadcasting, e.g. video/audio telecasting.

For the latter two operations, it is especially expedient to apply satellite networks as they provide global coverage and have star topology.

Main parameters of a few satellites networks are summarized in Tables 4.10.1 to 4.10.3.

Designation	IRIDIUM	GLOBALSTAR	THURAYA	New ICO
Orbit Satellite distance	LEO 780 km	LEO 1414 km	GEO 44°	MEO 10390 km
Number of orbital planes	6	8	-	2
Number of satellites	66+6	48+4	2	10+2
Beams/satellite	48	16	250-300	163
Bit rate kb/s	2,4/4,8	2,4/4,8/9,6	9,6 and GPS	4,8/38,4 144
Start of operation	Nov.1998	Oct.1999	May 2001	2003
Number of gateways	12	48	-	12 and ICONET
Voice channels per satellite	1100	2800	13750	4500
Note	regenerative	transparent	regenerative	regenerative

Table 4.10.1. Narrow-band MSSsystems

Designation	TELEDESIC	SkyBridge
Orbit	LEO 1400 km	LEO 1457 km
Number of orbital planes	12	8+8
Number of satellites	288	32+32
Beams/satellite	64 supercells 9 cells/supercell	45
Traffic	Multimedia/voice/data 2 Mb/s uplink 64 Mb/s downlink	multimedia/ data/video 60 Mb/s
Frequency	K_a-band	K_u-band
Inter Satellite Link	8 links/satellite	none
Start of operation	2004	2002

Table 4.10.2. Broadband LEO systems

Designation	ASTROLINK	SPACEWAY
Orbit	GEO	GEO
Number of satellites	9 (5 in orbit)	9
Beams/satellite	58	48
Traffic	Voice/data/video 16-9600 kb/s 6 Mb/s	multimedia/voice/data/ video 16-6000 kb/s
Frequency	K_a-band	K_a-band
On-board processing	yes, ATM-based	yes, ATM based
Start of operation	2005	2002

Table 4.10.3. Broadband GEO systems

4.10.2. Satellite information gathering and dissemination

In the examples in Sec. 4.10.1 on the applications of broadband ATM-based satellite systems, reference has been made on the feasibility of satellite information gathering and dissemination (multicast transmission, broadcast transmission). In this Section, a few specific applications will be discussed, summarizing the principles of Very Small Aperture Terminal (VSAT) systems, Satellite New Gathering (SNG) systems, and the role of satellites in the Internet network.

VSAT networks

VSAT networks are satellite private networks established by regional or global corporations for handling their data traffic, in some cases also voice traffic. Earlier, this requirement was met by the Intelsat Business Service (IBS), but in order to meet the increasing demand, Telecom General (USA) developed in 1979 their system designated VSAT, for serving hotel chains, banks, news agencies etc. At present, the

VSAT designation is no longer referring to the manufacturer, and is generally applied for business communication systems utilizing small (Ø 1.2 to2.4 m) antennas.

In VSAT networks, geostationary telecommunication satellites are applied, resulting in a single-path attenuation of 200 dB (e.g. in C-band) and a propagation time of at least 120 ms. The economically viable transmitter power of the terminals being at most a few Watts, direct communication between the terminals is not possible in lack of on-board processing on the satellite. This is why a central hub station is required, interconnecting the VSAT-s over the satellites via a star-shaped network. In Europe, most of the VSAT networks are operating in the Ku band, but on other continents, C-band networks are mainly used. Designations used in VSAT networks:

- outbound link: this is the path leading from the hub via the satellite to the VSAT, i.e. uplink + downlink.
- inbound link: this is the path leading from the VSAT via the satellite to the hub, i.e. uplink + downlink.
- one-way network: VSAT terminals are only receiving, over a point-to-multipoint communication system disseminating the information originating from the hub, e.g. news, data, TV programs, advertisments.
- Two-way network: VSAT terminals are both receiving and transmitting, e.g. bank transactions, remote printing, hotel booking, ticket sale, phone traffic, e-mail, medical data base, satellite news gathering (SNG).

Over the outbound link, TDM transmission is used, the VSAT terminals receiving data in their alotted time slots. In the TDM frame, synchronizing frames and addresses providing selective reception are used. Several access methods are applied in the inbound traffic, the most widely used being the TDM/TDMA system. However, over short durations of intensive traffic, one of the ALOHA accesses can also be used.

The VSAT terminal has two parts, the outdoor unit (ODU) and the indoor unit (IDU), interconnected by a cable having a length of less then 100 m. The indoor unit comprises the baseband interface to which the user terminal equipment is connected. Parts of the hub station: RF-unit (antenna, up- and down-converter, transmitter, receiver), indoor unit, comprising the baseband interface, to which the host computer (HC) is connected.

In communication systems operating over geostationary satellites, the long propagation time and high path attenuation, further their variations originating from

propagation phenomena, call for the use of error correcting coding and protocol conversion (emulation, spoofing) at the interface connecting the terminal equipment with the VSAT network. Protocol conversion is used at three layers (network layer, data layer, physical layer), but emulation may occasionally be required at the transport layer too.

A VSAT network with mesh topology can be established when signal processing is applied on board of the satellite. In this case, the VSAT stations are directly interconnected over the satellite, without the hub station. On board of the geostationary satellites operating in the Ka band (ACTS, SPACEWAY, ASTROLINK), signal processing is already applied. These satellites have higher transmitter power, and due to the Ka band, transmissions with larger bandwidth can be realized, thus upgrading the VSAT networks.

Satellite noise gathering (SNG)

At the beginning of the eighties, the increasing demand for TV news called for simultaneous picture and voice transmissions from remote areas in the case of unusual events (e.g. sites of the Gulf war, volcanic eruptions). These requirements are met by the private networks of TV companies, i.e. the satellite news gathering systems: over the satellite, reporter commentaries are relayed to the studio by terrestrial stations installed at the site.

There are two main types of SNG equipments, the portable (fly away) equipment, packed in containers and transported to the site by aircraft, and the equipment installed in a vehicle for transportation.

According to ITU-R Recommendations, SNG sysems operate in the C-band and the Ku-band. Simultaneously four SNG signals with digital modulation can be transmitted over a 36 MHz wide transponder. Systems operating with compressed transmissions are increasingly applied.

Internet over satellite

Satellite communication proved to be successful when terrestrial solutions became cumbersome because of the long period and high cost of the installation. In the eighties, the satellite market was enlarged by the large-scale proliferation of VSAT systems. However, these have been profitable for large companies only

because of the high price of terminals and of the transponder leasing: in 2001, appr. half a million VSAT terminals operated worldwide, two-third of them in the USA.

Recently, the broadcasting and dissemination of digital programs became a new field of satellite techniquies. In Sec. 4.7, this will be dealt with in details by summarizing the principle and composition of DVB-S and DVB-MS systems, and investigating the program broadcasting interactive systems.

Internet over satellite is the extension of the DVB standard for data transmission, i.e. the transmission of internet data traffic over the satellite digital platform [4.10.7].

There are two procedures for establishing the satellite internet network, the one-way access and the two-way access. According to the one-way access, the signal of the Internet data base (2 Mb/s to 40Mb/s) is transmitted by the hub station over the satellite to the user in the Ku-band. The satellite equipment of the user is applicable for reception only, and is thus an inexpensive device because the circiuits are identical with the mass-produced Small Office-Home Office (SOHO) devices. The return connection is established not over the satellite but over the existing terrestrial network, with a transmission rate of up to 64 kb/s.

According to the two-way access, the return connection is also established over the satellite, notably in the Ka band at a transmission speed of 144 kb/s to 2 Mb/s. The user equipment comprises a Ka-band transmitter and a Ku-band receiver, being thus more complex and more expensive. The composition and system parameters are given in the standard Digital Video Broadcasting Return Channel System (DVB-RCS), frequently referred to by the designation „ADSL-in-the-sky”.

These satellite Internet systems are operated by GEO satellites (Hot Bird, SESAT, ASTRA), but there are plans to operate also LEO systems (TELEDESIC, SKYBRIDGE) with the designation INTERNET-IN-THE-SKY)

References

[4.10.1] Almár-Both-Horváth-Szabó**:** Űrtan. SH Atlas. Springer Hungarica, 1996

[4.10.2] Nguyen, Hoang N.: Mobile Internet Provisioning in Satellite IP Networks: Mobility Management, Internetworking and Integration with Terrestial Networks. Proc.of 19th International Communications Satellite Systems Conference. 17-20. April 2001. Toulouse, France. Vol. 3. pp. 205-214.

[4.10.3] Boeke, Cynthia: Via Satellite 2001 Satellite Survey, Via Satellite. July 2001. pp. 24-30.

[4.10.4] Jamalipour, Abbas: Broad-Band Satellite Networks. The Global IT Bridge. Proc. IEEE. Vol. 89.No.1. January 2001. pp. 88-104.

[4.10.5] Walke, Bernhard H.: Mobile Radio Networks. Networking and Protocols. J. Wiley 1999.

[4.10.6] Pattan, Bruno: Satellite-Based Global Cellular Communications. McGraw Hill,1998.

[4.10.7] Blineau, J. Satellite Contribution to the Internet. Alcatel Telecommunications Review. 2001/4. pp.243-248

Translated by Tamás Sárkány dr.

4.11. The classical Internet

György Réthy PhD, author

Tamás Henk PhD, reviewer

The bases and current structure of IP networks are inseparable from the development of Internet. The development of the predecessor of Internet was initiated by the Advanced Research Projects Agency (ARPA) of the U.S. Department of Defence (hence called APRANET) to research of packet mode interworking of computer networks and interconnected university and research LANs. The protocols used in the network passed many developments phases until reached the specification of the even today (with amendments) used IP protocol family in 1981 (IP version 4). By 1983 the APRANET fully changed to TCP/IP protocols and in the same year the University of Berkeley issued the first Unix OS (BSD4.2) including (free of charge) the TCP/IP protocol family. Up to now, with the gradual spreading of IP based applications, the TCP/IP family become the prevailing protocol in enterprise data networks. Later a higher capacity backbone network still interconnecting research and university sites and called NSFNET (an abbreviation of National Science Foundation Network) replaced ARPANET. The capacity of NFSNET was expanded several times.

Commercial use of the Internet brought a definite change in the life and structure of Internet. The structure resulted is shown on Figure 4.11.1. Internet Service Providers (ISPs) are connected to each other via network access point(s) (NAPs), which are high capacity IP routers and also called shortly “peering”. NAPs are also connected to each other and their network forms the national and international backbone of Internet. Hungarian NAPs are jointly named BIX (Budapest Internet eXchange) and except exchanging IP traffic among Hungarian ISPs and bigger Intranets also provide international connectivity. Its structure is shown on Figure 4.11.2 (source: http://goliat.c3.hu/bix). More detailed information about BIX can be found at URL http://www.nic.hu/bix. The internal structures of ISP’s network consist of a router network that may be a simple or a multilevel hierarchical one depending on the size of the ISP/company. To these routers the users may be

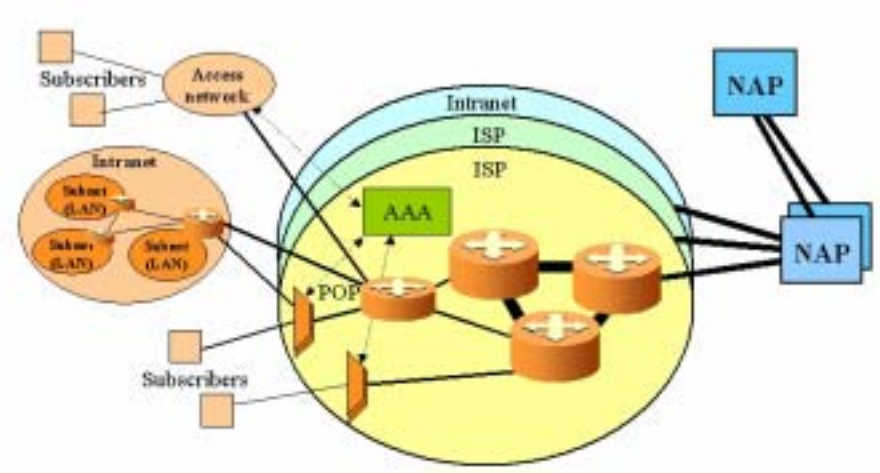


Figure 4.11.1 Simplified architecture of the Internet

connected in a permanent way via access networks like ADSL or CableTV or by switched means (analogue modems or ISDN) in a time-to-time manner via Internet dial-in equipment. Enterprise, institutional or other private networks, called in a general term Intranets, may be connected to ISP's routers directly or directly to a NAP (e.g. university and research networks, bigger enterprise Intranets).

The main building blocks of Internet are:

a) The TCP/IP protocols stack,
b) Network elements, routing and routing protocols,
c) Addressing and address resolution

In the chapters below we give an overview of each of them.

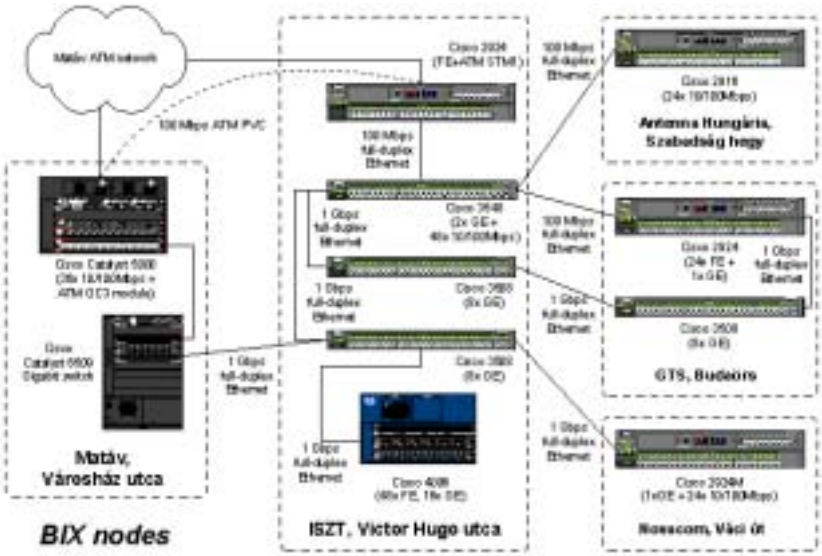


Figure 4.11.2 The structure of BIX

4.11.1. Internet protocols

The IP protocol architecture is shown on Figure 4.11.3. On the picture we also compare the IP and OSI layers (further difference is, that IP layer are less independent then layers in OSI). The structure of the IP header is shown on Figure 4.11.4. The most important tasks of the IP layer are network addressing of data blocks received from higher layers and their unacknowledged transport via the network in a connectionless manner (datagram mode). The transmission path is selected for every packet individually and the network does not guarantee a loss or delay performance (as we will see later, the change of the route is typical in case of failures and congestion in the network only). The maximum size of the IP packets is 64 Kbytes but majority of real networks is unable to transport of packets of this size. The actual packet size (supported on a hop) is called Maximum Transmission Unit (MTU) and determined by the network interface used (Ethernet, ATM etc.). For this reason blocks received from upper layers and exceeding the MTU are fragmented by the IP layer; on receiving IP packets containing these fragments at the remote site the IP layer restores them (fragmentation can be forbidden in the DF field of the IP header; in this case oversize packets are discarded). Further fragmentation and restoration may be necessary inside the network to transmit packets over networks

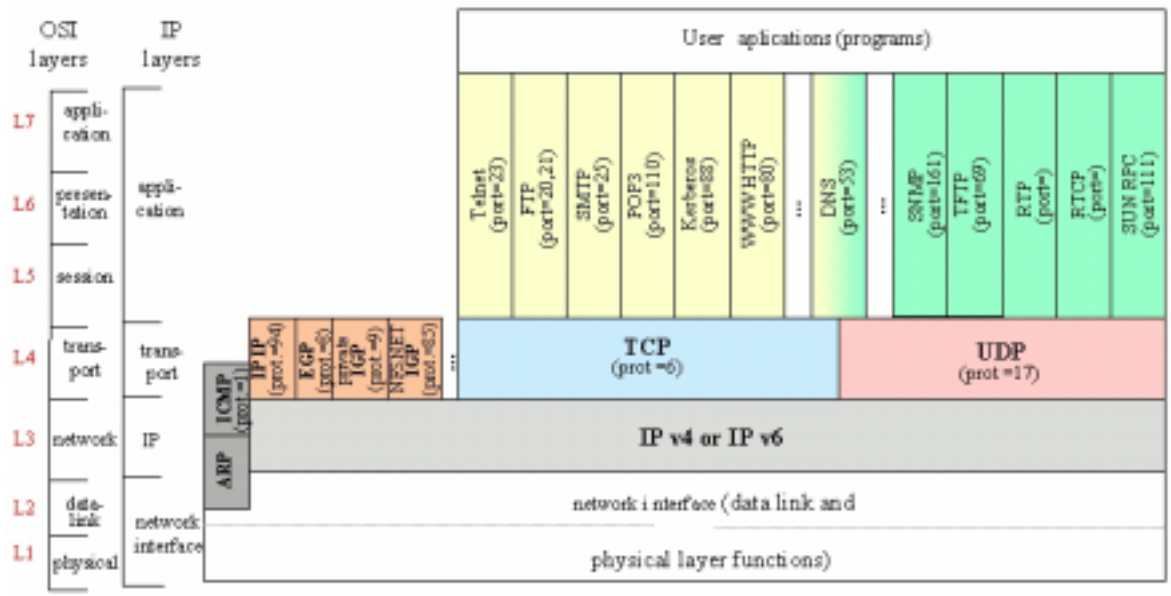


Figure 4.11.3. The IP protocol stack

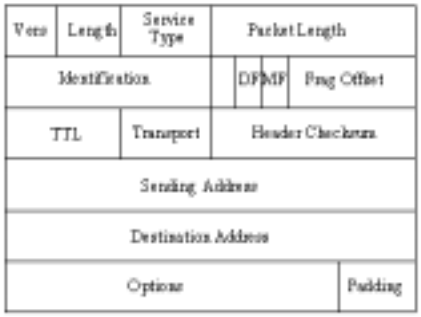


Figure 4.11.4 The structure of BIX

supporting an MTU less than was supported by previous networks.

The IP layer also avoids routing loop of packets. For this reason the time to live (TTL) field of the IP header is set typically between 15 and 30 seconds and each router decreases the value by as many seconds as the packet has spent at it but at least 1 second. Does the value of the field reach 0 the packet is discarded. The header (only) of the IP packet is protected from bit errors by a simple (and hence not too powerful) checksum, which, due to the change of (at least) the TTL field has to be re-calculated at the output of all routers. The ToS field of the header serves for the relative prioritisation of packets but majority of existing equipment does not support this feature. Option field following the IP header enables the exact or approximate pre-determination of the packet's route. If the route can not be followed, the packet is discarded.

From the point of view of addressing, the most important fields of the header are the source and destination IP addresses, which we will detail later. The 8-bit transport protocol field identifies the protocol, the protocol data unit (PDU) of which is carried by the IP packet. This protocol can be the IP control and management protocol (ICMP), a routing protocol or a transport layer protocol (handling user data). Except the most widespread TCP (Transmission Control Protocol) and UDP (User Datagram Protocol), the latter can also be one of the several vendor-specific proprietary protocols. Protocol identifiers used for IP encapsulation like „IP within IP" (value 94, see RFC 2003) used for mobile IP, minimal IP encapsulation (value 55, see RFC 2004), IP encapsulation header (value 98, see RFC 2004) or encapsulation of IPv6 packets by an IPv4 header (RFC 2893) shall be highlighted. These

codepoints enables the transport of an another IP packet – even in a recursive way - in the data field of an IP packet. Such transport is called an IP tunnel and is extensively used for different purposes. Currently there is about 100 transport protocol codepoints allocated.

From the functional point of view ICMP (Internet Control Message Protocol, RFC 792) is a network (IP) layer protocol, while during transmission it is handled like a transport layer protocol and ICMP packets are carried embedded in the data field of IP packets. Its functions are the indication to the sender of discarding a packet within the network (due to TTL reaching zero, unknown destination address, a syntactical error etc.), changing of the route (e.g. optimisation of the path between gateways) and assisting traffic control (e.g. source quench). For example, the popular ping program also uses the Echo Request and Echo Reply procedures of ICMP.

The UDP just adds its own header to the user data packet and passes its to the IP layer. Therefore an UDP session inherits the properties of the IP session: may have loss and unacknowledged. The header contains the source and destination port numbers, a checksum and a length indicator. Interesting that the UDP checksum is calculated from the source and destination IP addresses in addition to the UDP header and data fields. This is an additional protection against misdelivery of IP packets due to undiscovered bit errors in the IP header.

The function of TCP is the reliable and acknowledged transfer of duplex data streams over IP. TCP segments the data block received from the application layer to the size required by MTU (in contrast with UDP, which expects data blocks of the size consistent with the MTU from the application layer). TCP fulfils several subtasks, as traffic control, multiplexing, fast retransmission (retransmission of lost packets before timer expiry), congestion avoidance. The TCP transmission is connection oriented; data is transferred between session establishment and release. The window size is adaptive: at session establishment it is 1 and increasing exponentially until experiences packet loss. Then the window size is decreased to the half of the actual size and tuned linearly close to the throughput limit of the given path . The entity in the host, which initiates the session establishment is called active, while the entity connecting to the well-known port number when an application is started (i.e. an FTP server) and waiting for the incoming session establishment is called passive.

Both UDP and TCP headers carry the source and destination port numbers that enables the transport protocols to identify the application using their services. The host dynamically to its active applications allocates port numbers between 1023 and 65535; numbers below 1023 are the well-known port numbers. These are allocated to application types by the Internet community (RFC 1700, newest numbers are registered by IANA) and server applications are waiting for an incoming TCP sessions or UDP packets on these numbers. Well-known numbers are allocated to “standard” Internet applications as HTTP, FTP, and SMTP etc. as well as to proprietary protocols used within enterprise networks. The host will initiate session to the server application at the well-known port number. Port numbers alone may not identify a given session in all cases. The reason for this is that the server may serve more than one host simultaneously, which may use the same source port number (see Figure 4.11.5). In this case the server is able to identify the session with the complex entirety of source and destination IP addresses and port numbers only. These four identifiers together with the type of the protocol are called the socket number.

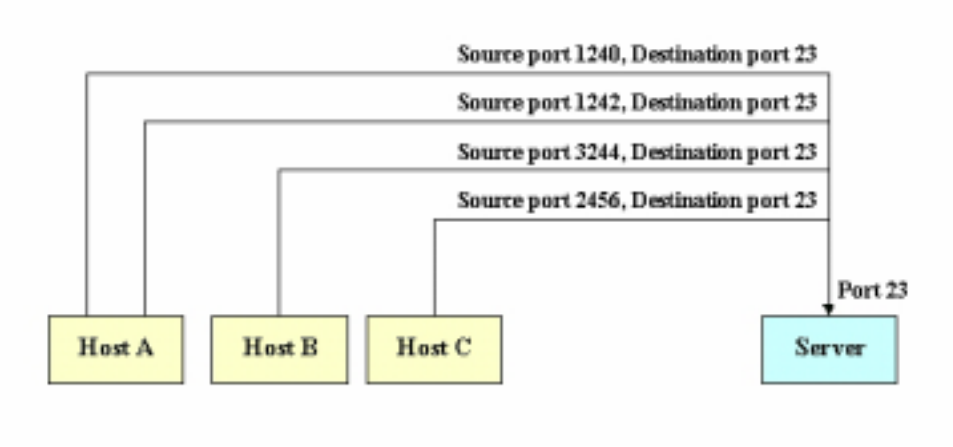


Figure 4.11.5 Port multiplexing

4.11.2. Internet addressing

The currently used IP v4 addresses are 32 bits long and consist of a network address and a local address part. Addresses are categorised into 5 classes as shown on Figure 4.11.6, from which individual networks as Intranets, ISP’s networks may get an A, B or C class address according to their size (local address space required).

IP addresses are written in a dotted quad notation, by four decimal numbers separated by dot (e.g. 147.23.89.2). Therefore a digit covers 8 bits in the address and its value is between 0 and 255. Parts of the address, which is all 0s or all 1s, have special sense: the former means “this”, all 1 identifies broadcast addresses. Therefore “0.0.0.0” identifies the own host, “0.0” written in place of the network address part of a B class address identifies the own network, “255.255” in place of the local address part of the same B class address means all hosts within the network. The „255.255.255.255” broadcast address is used for discovery of servers, whose address is unknown (e.g. DHCP server, mobile IP agents) and “127.0.0.0” identifies a loop within the own machine just before the physical interface.

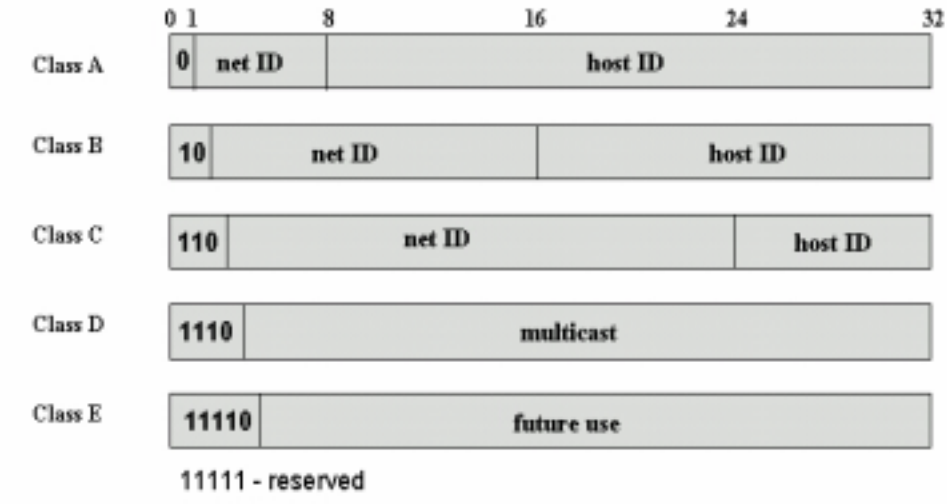


Figure 4.11.6 Structure of IPv4 addresses

IP addresses may be public or private. Public addresses are the permanent and known addresses of registered networks allocated by the Internet community (and registered by IANA). In all classes there are network addresses allocated for private use (e.g. 10.0.0.0/24, 192.168.0.0/16). Any network freely can use these internally but they must not be used for the traffic between networks (on the Internet). In these cases gateways shall translate the internal (private) address to one of the public addresses allocated to the given network. This is called Network Address Translation (NAT).

Bigger networks are usually decomposed to several sub-networks (or shortly subnets), e.g. the Intranet of a company to subnets according to physical sites (typically for routing, traffic optimisation and for easier manageability). As the IPv4

address space is limited, it would be a wasting to allocate their own (public) network addresses to smaller subnets. Therefore the local address part (of the public address) of the given Intranet's address space is further separated into subnet and local address parts. As this separation is not seen from the IP address itself (unlike the class of a public address), a 32-bit subnet mask is also distributed to the hosts within the subnet. In the mask "1" identifies bits belonging to the subnet address and "0" identify bits belonging to the local address of the given subnet. The network and subnet parts of the address are also called an address prefix.

As part of the IP routing the host, before sending a packet fits the mask with the destination IP address and its own IP address. If the two results equal, the host resides within the subnet and the packet to be sent to the LAN address of the given host. If the two results differ, the wanted host is outside the subnet and the packet shall be sent to the LAN address of the gateway.

A similar problem exists for big networks. There are networks which can not fit into a single C class address range can to obtain a B class network address, which is of huge shortage anyway. These networks are getting several C class network addresses what would lead to the unnecessary increase of routing records and routing tables (see section 4.11.5) on the backbone. Intranets owning a group of network addresses are also called supernets. Therefore the masking technique described above is used to form address groups used at routing information distribution. This is called Classless Inter-domain Routing (CIDR, see RFC 1518 and RFC 1520). Notice the difference between the use of the mask in subnets and in supernets. It is used in subnets to chunk the network address space to smaller subnet address spaces, hence the number of "1" bits is more than the length of the network address. In case of CIDR the use is the opposite – to group network addresses - and the number of "1" bits is less than the length of the network address.

4.11.3. IP v6

Partly due to size restrictions of this section, partly before it is not yet used on the Internet, only the two most significant features of IP v6 are mentioned here. IP v6 is described in more detail in subchapter 3.7. In contrast to the 32-bit address length of IP v4 (which, in spite of all provisional measures as dynamic address allocation,

subnet addressing etc. is expected to become exhausted within a few years) the length of IP v6 addresses is 128 bits and support three built-in addressing mechanisms: unicast, multicast and anycast. The IP v6 header contains priority and QoS stream identifier fields, allowing to eliminate for example one of the most significant drawbacks of Intserv (see subchapter 4.12), and the packet-level classification. IP v6 is capable of avoiding fragmentation on singular network sections within the network and supports authentication and encryption. The header structure of it is simpler then in IP v4; all fields are the multiple of 4 bits and not protected by a checksum. The first header contains the most important information only. All other additional information influencing routing, fragmentation and other functions are contained in one or more subsequent headers. All these measures result quicker processability of IP v6 packets. It shall be mentioned, that IP headers contain a version identifier field, therefore in spite of the different header structure, the same network can carry IP v6 and IP v4 packets simultaneously.

4.11.4. Internet routing

Routing it one of the most important (or dare to say the most important) function of IP networks. It resides in hosts and servers as well as in network equipment like routers and gateways. We have to discuss shortly the difference between routers and gateways. As a simplified definition we can say, that routing is a network layer packet forwarding function, which determines the output interface for the given packet, while gateways cover the functionality by which a subnet (e.g. a LAN, inside which direct sending is used) is interfaced with other subnets and the Internet. In other words the gateway is a router with additional functions and further on, if no need to emphasis gateway functions, we will use the more generic “router” terminology for both routers and gateways.

The routing functions consist of two components: handling of routing information and selecting the next hop, where the processed packet to be sent. The routing function in hosts has been described in section 4.11.3. This section summarises functions in routers. Selecting the next hop is done by a rather simplified way on the Internet called the best effort service. This means that routers serve all packets in the order of receipt and forwards them to the output believed to be optimal. In other words routers handle a single QoS class. As a result the network

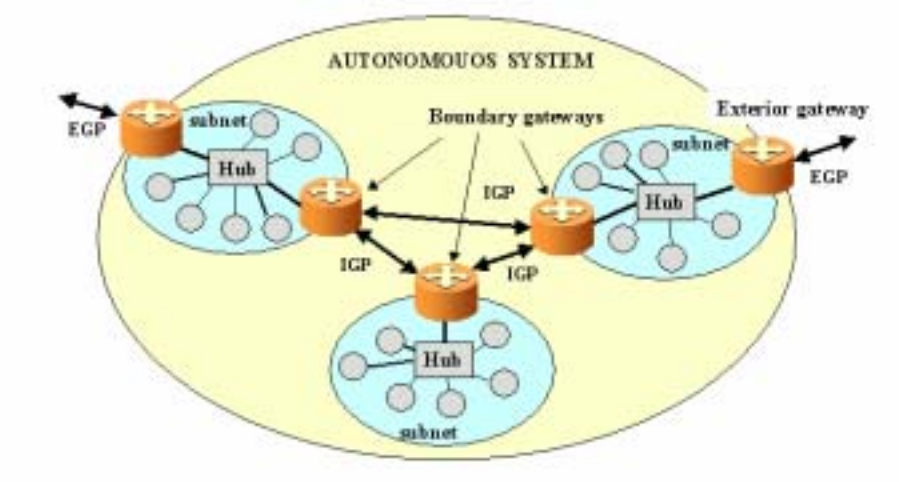


Figure 4.11.7 Structure of network routing

gives no bandwidth or delay guarantee, packets may be lost, their order within a stream may be inverted and packets may multiply within the network. The IP layer does not handle these cases; it has to be considered in higher layers.

All routers have their own routing table containing static and dynamic records. Dynamic means here, that time-to-time or in case of changes in network capabilities (e.g. interface failures, congestion) routers are sending routing information (IP addresses and related costs) to all neighbouring routers, they say advertising it. This advertisement is done by routing protocols. Structure of the Internet from the point of view of routing is shown on Figure 4.11.7.

IGP (Interior Getaway Protocol) and EGP (Exterior Gateway Protocol) identifies the role of the routing protocols in the Internet architecture, therefore is a functional distinction. If for example, a protocol designed to fulfil an EGP role is used to route traffic within an autonomous system (AS), it would become an IGP (it is not true vice versa, as a rule an IGP can not be used in the EGP role). From the viewpoint of Internet an autonomous system (AS) is the network part, which belongs to a single administrative domain, i.e. an Intranet or an ISP’s network. All subnets have at least one gateway and an IGP (type) routing protocol is used between subnets gateways, i.e. within an AS. An EGP (type) protocol is used to interconnect exterior gateways of ASs, so in the Internet backbone.

Though several IGP protocols exist, the most widespread are the RIP-2, IS-IS and OSPF protocols. RIP-2 (Routing Information Protocol version 2, RFC 1723) has corrected several problems of the first version. As RIP-2 is not a new protocol (accepted in 1994), further on in this section we will describe RIP-2 without mentioning the differences. The cost belonging to a known IP address is a simple distance vector measuring the number of hops between the given router and the destination network. The distance of the directly connected network or router is 1; the maximal distance (non-reachable network) is 16. All routers advertising its whole routing table at start up, time-to-time (typically every 30 sec), on request or in case of any change in the table (change of an own link state, new information from the neighbours etc.). For all known IP network addresses the router will choose the neighbour with the smallest cost as the output direction and will forward all packets with the given address to this neighbour. This simplicity is paid by a higher convergence time (the time needed after a route change until all router will have a stable updated routing table) and a network size limitation. The problem is caused by the “counting to infinity” phenomenon. At the example shown on Figure 4.11.8 a) to the address prefix of the destination net records in different routers will be as shown in Table 4.11.1. Notice that R1 and R2 mutually has reference to each other but this route is currently not used. If the link between R3 and the destination networks fails (see figure b)) it advertises its new table in which the destination network is shown as via R1 and R2 on distance 4. R1 and R2 receives this information approximately at the same time, they update they tables, showing now the smallest cost of the destination network via each other at distance 3 (in both routers). They send this information to everyone and in the next step – mutually receiving updated routing information from each other adds to the new value received the distance between R1

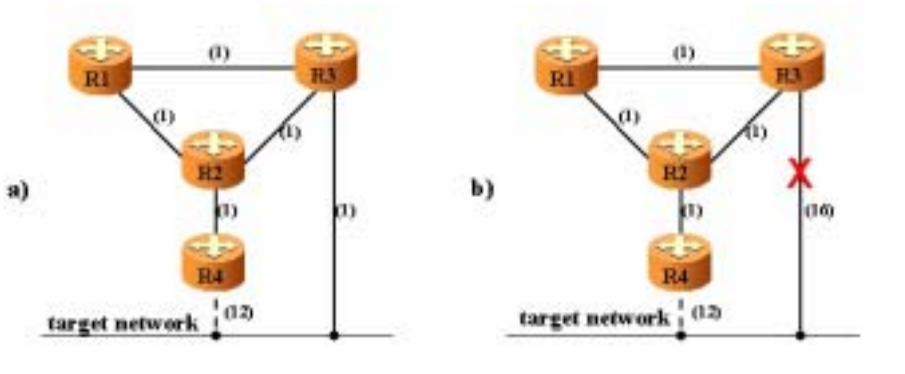


Figure 4.11.8 Counting to infinity

and R2 (1), updates the table again to show the destination network via each other at the distance 4 etc. So they will increase the cost by one in each information exchange until the route via R4 becomes the cheapest one. Now, the procedure stops showing distances 13 in R2 and 14 in R1 and R3. If this high-cost route and the limit of the distance value did not exist, R1, R2 and R3 would step-by-step reach overflow of their counters.

This instability is decreased by several methods. At the split horizon methods the routing record of the prefix for which the cost is changed is not sent back to the neighbour router the change has been learnt from. At the split horizon with poison reverse method the record is sent back but with the cost 16. The benefit of the latter solution is that a “non-reachable” record will be mutually written in the first exchange of routing information excluding the forming of a loop. The drawback is, that routing information sent is becoming bigger what may cause performance degradation on low bandwidth links. Maximising the hop count in 16 causes that routes can not distinguish different routes for addresses more than 16 hops away. RIP also may cause congestion in some cases, because neither the bandwidth of links nor he throughput of routers taken into account. Hence for realtime traffic it is unusable or can be used with serious restrictions only. But in middle-sized networks without significant capacity bottlenecks and carrying mainly data RIP performs good and its operation is simple.

For handling of huge networks OSPF is more suitable (which has much more advantages than simple this), which in contrast to RIP, is a link-state protocol. Its basic principle is very simple but - as a rule – its implementation becomes quite

Records in router…	Next hop	Distance	Status
R1	R3 R2 R4	2 3 14	chosen route
R2	R3 R1 R4	2 3 13	chosen route
R3	Directly R2 R1	1 4 4	chosen route
R4	R2 Another route	3 12	chosen route

Table 4.11.1. Routing table records to Figure 4.11.8 a)

complex. Each router requests by “hello” packets identifiers of all of its neighbours and allocates a cost reflecting the state of the link (bandwidth, load etc.) to the given route. From the data gathered in this way a link state packet (LSP) is composed and the LSP is advertised using the so called flooding technique to all other routers within the AS; namely when a router receives an LSP it will distribute this LSP to all its neighbours, except the one it received the LSP from. LSPs are numbered; therefore new information can be distinguished from an elder still “wandering” one. Routers calculate the topological table of the network from all LSPs received. A router sends an LSP in every 30 minutes, if discovers a new neighbour or the state of the link to one of the neighbours has changed (down, up, congestion etc.). Further advantages of OSPF that it allows traffic sharing between links and is capable of handling multiple QoS classes by allocating different costs to all supported classes.

Examples of EGP protocols are EGP (RFC 827, RFC 888, RFC 904) and BGP-4 (Border Gateway Protocol version 4, RFC 1771). EGP is still a recommended protocol but de-facto become historical. Hence below properties of BGP-4 are summarised.

Most important features of BGP-4 are:

a.) Supports CIDR depicted in section 4.11.3; e.g. the subnet/supernet mask is attached to each network address being announced. This allows advertising a group of network addresses as a single address prefix, which will be recorded by other BGP gateways as a single destination address.

b.) It is a distance vector routing protocol but in contrast to RIP, more factors are considered in finding the minimal cost route. It enables to avoid forming of routing loops experienced with RIP.

c.) It supports routing policies. This function is not part of the protocol but of the implementation, RFC 1772 recommends a (minimum) set of policies, which enable AS administrators to define: which address prefixes may be advertised to which another AS; which another AS may use the network for transit traffic. A weight may be allocated to each external AS which will be considered by the exterior border gateway together with the cost of different routes when choosing an optimal path. By setting the weight to infinite some ASs can be totally excluded from routing.

Three different types of ASs distinguished by BGP: a) stub AS, which has a single inter-AS connection; b) multihomed AS, which has connections to multiple ASs but refuses to carry transit traffic; c) transit AS, which has connections to multiple ASs and carries both local and transit traffic. The type of an AS is determined by the

administrator of the AS domain and enforces via the routing policy put into its border gateways.

BGP-4 uses TCP port 179. The BGP gateway will initiate TCP connection to its neighbour at this port number. After the TCP connection is established, the initiator sends its identifiers (AS number and BGP identifier) in an OPEN message including also the requested maximum hold timer value (recommended value is 90 s) and optionally authentication information. Neighbours are maintaining the connection by sending BGP messages periodically: if no routing information to be sent for a longer time (typically for 30 s) a KEEPALIVE message will maintain the connection. The first KEEPALIVE message shall be sent as a response to the OPEN message. If the hold timer expires in a BGP gateway it will send an error notification to its neighbour and disconnect the TCP connection.

Neighbours are advertising routing information by UPDATE messages. A single update message is able to advertise a single feasible (reachable) route and withdraw several infeasible routes. Unlike from RIP, BGP-4 informs not only about the network prefix of the advertised net but also about the path leading to it. This feature not only allows determining the cost of a route, but enables gateways to avoid routing loops and the phenomenon of counting to infinity experienced with RIP. At the time of establishing the TCP connection BGP routers will send UPDATE message for each address prefix to be advertised. Later they will send only updates to the routing information distributed at the beginning.

BGP gateways shall send a NOTIFICATION message in case an error occurs like BGP message header error, OPEN message error, UPDATE message error, hold timer expiry etc.

4.11.5. Mobil Internet

In all the above discussions we implicitly assumed that hosts are connected to the net at permanent attachment points (from the point of view of this section we will consider a host dialling into an ISP’s network to be connected at a permanent attachment point). Mobile IP represents a different capability. The word “mobile” has narrower meaning here than we used to in mobile telecommunication (like GSM) networks. It covers terminal mobility only, i.e. the mobile host, for example a

notebook, on connecting to the Internet in different locations, also remains accessible on its home IP address (for this reason also called a nomadic IP or nomadity). However, this can be taken as on eof the forerunners of future mobile IP communication. The mobile IP solution is described in RFC 2002 and RFC 2003. Its basic principles are shown on Figure 4.11.9.

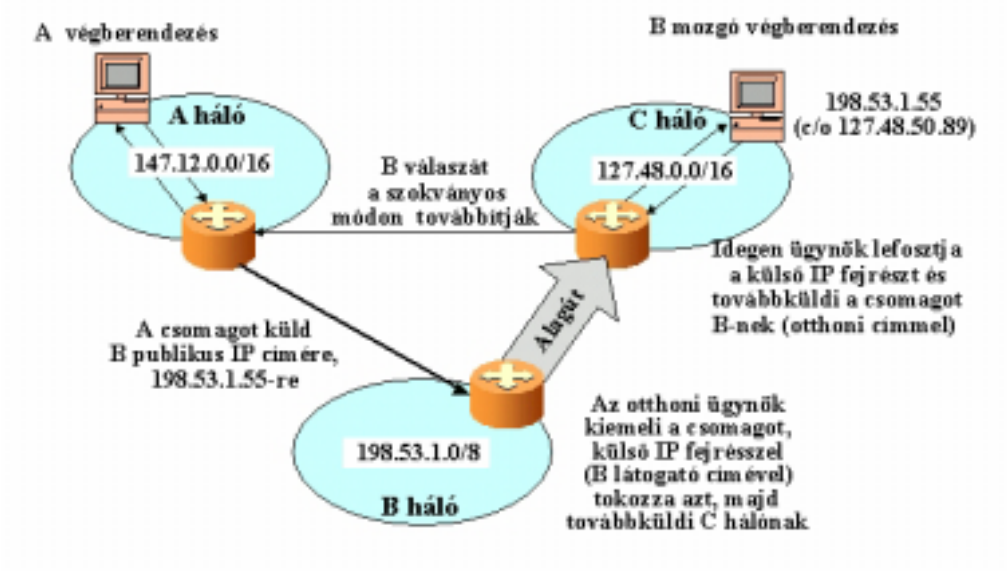


Figure 4.11.9 Basic principles of mobile IP

The mobile IP architecture has at least one new network element; the home agent and naturally the host also shall support mobile IP functions. The home agent advertises himself on the home network by special ICMP messages but the mobile host also can inquire if there was any on its home subnet. When finds one, the host stores the IP address of it. When the mobile host is connected to a foreign network, it tries to find a foreign agent, who also advertises himself by ICMP messages. If finds one, it registers itself at the foreign agent and asks a care-of-address, what is one of the public addresses of the foreign network. The mobile host then registers the care-of-address at its home agent who, from now on, knows where to forward packets for the mobile host to. If an IP packet arrives at the home subnet for the mobile host home address, the home agent intercepts this packet and forwards it to the care-of-address of the host via a tunnel (see section 4.11.2). The source and destination addresses of the outer header will be the IP addresses of the home and the foreign agents consequently. The foreign agent terminates the tunnel by stripping the outer header and sends the packet to the mobile host leaving the inner header with the

home address. The existence of the foreign agent is optional. For mobile IP communication it is enough, if the DHCP server on the foreign net has a dedicated address space for (foreign) mobile hosts. In this case the above procedure slightly changes, and the mobile host will terminate the tunnel, therefore the packet passes the foreign network using the care-of-address in the outer header.

For sending the mobile host is using its home address as the source address and packets are routed to the destination by normal routing procedures. When the mobile host returns to its home network it will de-register itself from the home agent and goes on functioning as a usual host. It worth to mention, that IP v6 supports mobility, due to address autoconfiguration and neighbour-discovery there is no need for a foreign agent at all.

References

The joint literure of sections 4.11 and 4.12 is given at the end of section 4.12.

4.12. Realtime IP networks

György Réthy PhD, author

Tamás Henk PhD, reviewer

4.12.1. Directions of further developments

As described in subchapter 4.11, Internet is designed to be a robust data communication network and up to now has preserved this attribute. Rapidly increasing number of Internet users and bandwidth raised the question of providing conventional speech services over IP based networks (VoIP) could be cheaper than over existing circuit switched ones. Quality requirements of such a service are specified by ETSI standard TS 101 329-2[36] [4.12.38]. Similar work has been started in ITU-T (see draft Recommendation Y.1541) and in IETF (RFC 2330). TS 101 329-2 specifies the quality classes and parameters shown in Table 4.12.1.

	3 (WIDEBAND)	2 (NARROWBAND)			1 (BEST EFFORT)
		2H (HIGH)	2M (MEDIUM)	2A (ACCEPTABLE)	
Relative Speech Quality (one way, non interactive speech quality)	Better than G.711	Equivalent or better than ITU-T Recommen-dation G.726 at 32 kbit/s	Equivalent or better than GSM-FR	Not defined	Not defined
Overall Transmission Quality Rating (R)	N/A	> 80	> 70	> 50	> 50
End-to-end Delay	< 100 ms	< 100 ms	< 150 ms	< 400 ms	< 400 ms
NOTE: The "R" and delay for best effort class are target values.					

Table 4.12.1 Tiphon QoS classes and their basic parameters

The relation between overall transmission quality rating (R) and user perception of quality is defined in ITU-T Recommendation G.109 [4.12.48]. Table 4.12.2 is extracted from this recommendation and shows this mapping.

[36] While ITU-T Recommendations and IETF RFCs and STDs are not legal standards, ETSI deliverables of the type ETS, EN and TS are de-jure standards in EC and associated members' countries including Hungary.

Overall Transmission Quality Rating	$90 \leq R < 100$	$80 \leq R < 90$	$70 \leq R < 80$	$60 \leq R < 70$	$50 \leq R < 60$
User's satisfaction	Very satisfied	Satisfied	Some users dissatisfied	Many users dissatisfied	Nearly all users dissatisfied

Table 4.12.2 Relation of the R factor and user satisfaction

Today's Internet is not fulfilling these requirements for it was not designed for it (in fact even requirements of the best effort class are unsatisfied as they are not guaranteed networkwide). Consequently, in IP networks technical development is needed to be able to support real time services (categorisation of IP based services see in subchapter 3.7). Two views are prevailing in answering the how question. One of them says, that over-allocating links bandwidth and router capacities can assure the required quality. According to the another, this way would be too costly as the most expensive part of Internet is the backbone and it would be economically unjustified to allocate real time quality to data communication, which is expected to be the majority of the backbone traffic. Besides, there always will be low-bandwidth (local) sections and their users can not be excluded from real time services. The latter argument especially true for mobile networks, even in the case of the emerging 3rd generation mobile family.

There are two mean technical areas of the development: providing the signalling capabilities for telephony and multimedia calls and assuring the real time quality of the established IP data streams. For the first area, examples are the H.323 and BICC architectures of the ITU-T and SIP/SDP protocols of the IETF. These are not only new protocols in the network but changes the whole architecture of the network. For the latter examples are IntServ, DiffServ and MPLS.

Below we give an overview of all of them. It has to be emphasise, that IntServ, DiffServ and MPLS does not exclude each other. All of them have pros and contras, therefor on long term mixed solutions can be expected (like IntServ+DiffServ, MPLS+DiffServ etc.).

4.12.2. Call control solutions

Today two major solutions exist to handle calls in VoIP and multimedia IP communication: H.323 and SIP/SDP. This section just gives a short summary as they are described in more detail in subchapter 3.5.

The H.323 multimedia conference system inherited its name from ITU-T Recommendation H.323, which is a frame document and describes the system on a functional level. Other ITU-T Recommendations specify different protocols used in the system. It supports the set-up and release of IP based multimedia calls, personal mobility (the user registers its location in the network and incoming calls follow him to the last host registered) and numerous supplementary services, which are almost identical to supplementary services common in PSTN/ISDN networks.

SIP/SDP protocols has been specified by IETF and are an alternative of H.323. SIP (Session Initiation Protocol, RFC 2543) is a call (session) control protocol, sets up, modifies and releases two or multiparty single or multimedia sessions. A session is not necessarily a voice or multimedia call, one can joining a point-multipoint session like a multimedia presentation. SIP also supports personal mobility but the few supplementary services supported (e.g. CLI) is incompatible with PSTN/ISDN supplementary services. SIP and SDP were originally designed for the Internet but 3GPP has chosen them as the call control protocol of the multimedia subsystem of 3rd generation mobile networks.

4.12.3. Integrated Services

IntServ (IS – Integrated Services, RFC 1633, RFC 1727, RFC 2210) is based on the recognition that required quality can be guaranteed by handling real time traffic separated from data traffic and by reserving necessary resources in advance. Intserv has tree mean components: admission control, resource reservation and traffic control. Admission control and resource reservation are two separate procedures but carried out simultaneously before any user data transmission. While resource reservation carries information along the network on the requested bandwidth and QoS between network elements (including hosts) and designates the route for the traffic, admission control is acting independently in all routers, and decide if the new resource request is acceptable without endangering the QoS

guaranteed for the existing traffic or not. The traffic control acts during the data transmission phase, controls that the data stream complies with the traffic descriptors given in the resource reservation phase and polices the stream if needed. For resource reservation the RSVP protocol (RFC 2205) is used. Basic principles of IntServ are given below:

RSVP procedures is initiated by the source by sending a Path message (see Figure 4.12.1) which does not reserve any resource but designates the route for RSVP messages and for the traffic and carries source traffic descriptors. RSVP messages has to be sent continuously during the traffic transmission phase. This enables to change the path if the required bandwidth and QoS can not be provided on the designated route due to network failure reasons. On receiving the Path message the destination host compares the source traffic descriptors with its own capabilities and decides what resources should be reserved within the network for the service. Then sends the Resv message, which includes the requested bandwidth and QoS parameters. All routers decide the accommodation or rejection of the request independently. If the request is accepted, the router records the admitted bandwidth (Tspec) and QoS values (Rspec), to be used later for traffic control, and the source and destination IP addresses and port numbers (all together is called filterspec), allowing to identify the related packet stream and allocates a QoS identifier to the stream. If the admission is unsuccessful, the router rejects the request with a ResvErr message sent to the destination host (which may retry the reservation for a lower bandwidth).

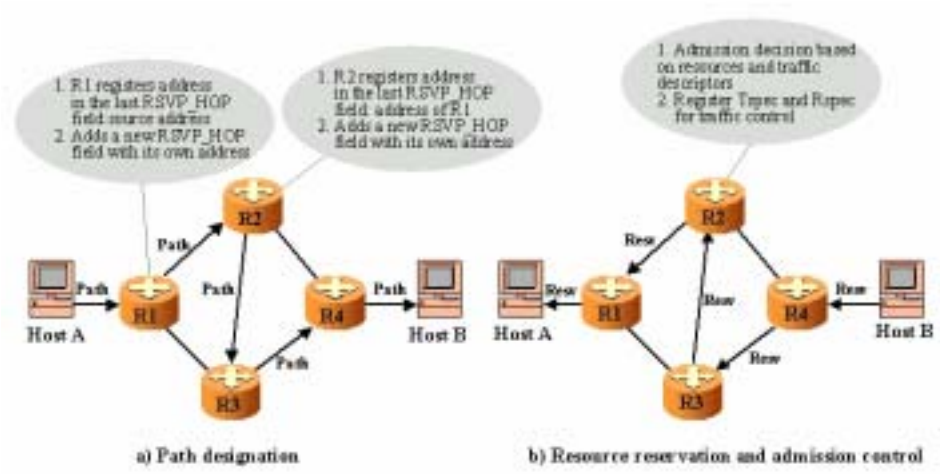


Figure 4.12.1 Principle of RSVP

When a packet arrives to an IntServ enabled router, first the classifier will take action, and based on the recorded filterspecs decides, if the given stream has a reservation or not. If it has, the packet is handed to the scheduler, which performs traffic control and policing. The policing technique used is the adaptation to variable packet sizes of the leaky bucket method proven in ATM networks and is called “token bucket filter”. Accordingly the traffic descriptors are the mean bitrate, burst size and the minimal and maximal packet sizes. The size of an individual packet may be less then the minimal packet size but the scheduler will take them into account at the minimal packet size (therefore they decrease the guaranteed bandwidth).

As can be seen the reservation of resources is done in the backward direction, i.e. from the sink to the source and always unidirectional. For bi-directional traffic (symmetric or asymmetric) both ends shall initiate reservation procedures in parallel (they will be handled independently by the network).

Currently IntServ differentiates two traffic classes: the Controlled load service, which assures the delay of a low-traffic router at the speed of the guaranteed bandwidth, and the Guaranteed Service which guarantees a maximised delay value. These services are described in more detail in section 3.7.5.

One of the most significant drawbacks of IntServ is the per packet classification of the incoming stream. This causes a significant processor and memory load in bigger networks (what makes them more expensive) and the delay produced by this processing is a quality decreasing factor itself. Except this, the renewal of the reservation also triggers some additional load.

4.12.4. Differentiated Services

DiffServ (DS – Differentiated Services, RFC 2475) intends to eliminate drawbacks of IntServ. DiffServ has two determining components:

a) categorisation of streams into QoS classes and identify the class in the packet header,

b) specification of the router behaviour which depends on the QoS class of the processed packet.

DiffServ redefines the ToS (Type of Service) field of the IP v4 header (which is mostly unused) and renames it to the DS field: 15 subclasses within 4 different QoS

classes are specified. All subclasses has an allocated codepoint (the DSCP) in the first 6 bits of the DS field, the two remaining bits are unused.

The class selector class aims to preserve compatibility with the precedence (the first three) bits of the IP v4 ToS field, therefore these bits are unused or used to provide a default behaviour which corresponds to the Internet best effort service. He first really new DiffServ class is the AF (Assured Forwarding, RFC 2597) class. Network operators may allocate in their routers different relative bandwidth and buffer sizes to the four specified AF classes (AF1..AF4) according to their QoS policy. Packets within any AF stream shall be categorised into one of the three different drop precedence subclasses, which determines the relative importance of the packet within the stream. In case of a traffic congestion packets of lower drop priority are discarded first. Therefore all AF subclasses uses up 12 DSCPs. An important remark, that the four AF classes are independent from each other, congestion within one of them shall not lead to quality degradation in the others. The EF (Expedited Forwarding, RFC 2598) class is defined as a low-loss, low-delay and low-jitter class and is meant for realtime traffic.

The another important component of DiffServ is the router behaviour that shall ensure transport characteristics like delay, packet loss and jitter described for the QoS class indicated in the packet header. However QoS characteristics shall be granted for the aggregated traffic of a given class (better to say DSCP) which is called a behaviour aggregate (BA) and not for individual streams. Each BA has a dedicated per hop behaviour (PHB) which describes packet handling procedures for the given BA. For example a packet of the EF class gets an absolute precedence over all other traffic and when arrives at a router it will be served immediately and will leave the router first (supposing no other packet of the EF class being processed). As no absolute QoS parameters (guaranteed bandwidth, packet loss, delay, and jitter) are specified, actually impossible to design a network providing QoS guarantees in a vendor-independent way. One of the most important aspects of DiffServ is the technique used within routers to assure expected PHBs. This is described in more detail in section 3.7.6.

Further drawback of DiffServ is that no source traffic descriptors are specified and policing is unfair as applied to behaviour aggregates. Consequently when one of

the users exceeds the traffic contract (see SLA in subchapter 3.7) the network will punish not only the user committing it but also all others.

Table 4.12.3 compares advantages and disadvantages of DiffServ and IntServ. Basically it is true in both cases that all network components shall support the relevant QoS technique. However while for the use of IntServ there must be at least one fully IntServ enabled path along the network, DiffServ can be used even if only part of equipment supports it (but the end-to-end quality can not be guaranteed).

IntServ and DiffServ are solutions not excluding each other. On the contrary, their joint use may give an optimal result. Due to size limitations we can not detail it here but its description is given in RFC 2998.

	Advantages	**Disadvantages**
IntServ	explicit admission control, bandwidth guarantee, source traffic known, per source policing, designated route for a stream	no explicit QoS indication, big processor and memory need, one QoS class per stream, soft reservation (needs renewal)
DiffServ	explicit QoS indication in the packet header, defined QoS classes, router behaviour is defined (PHB), per packet QoS (substreams with different QoS within a stream)	no admission control, no guarantee (congestion possible), per group (BA) policing, per group (BA) traffic conditioning, no source traffic description

Table 4.12.3 Comparison of Intserv and Diffserv

4.12.5. Multiprotocol Label Switching

The Multiprotocol Label Switching (MPLS) was primarily developed to provide efficient IP VPN services but it also can be used to offer quality of service. At first because it allows quicker switching, at second it classifies packets only once at the ingress router and not in all routers. The essence of MPLS is, that forwarding of a packet is not based on the IP header but on a fixed-length label attached before the header (see Figure 4.12.2). For this reason the MPLS part of the network shall be separated, it can not be mixed with conventional IP routers but so to say there shall be a closed MPLS “cloud”. MPLS equipment located inside the cloud are called label switch routers (LSRs), equipment located at the periphery of the cloud - which are in contact with the non-MPLS enabled world - are called label edge routers (LERs). When we do not need this functional distinction between MPLS equipment they will

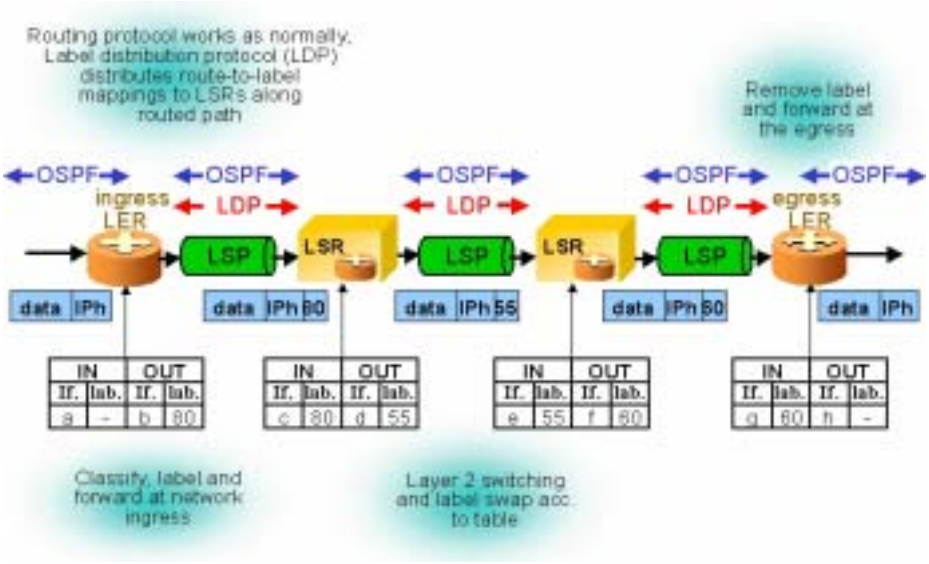


Figure 4.12.2 Principle of MPLS

be simply marked as LRs. The transmission path between LRs is called a label switched path (LSP). MPLS is composed of the control and the forwarding functional groups.

Control functions consist of the same IP routing protocol (IGP or EGP) as in non-MPLS IP networks but additionally, in parallel with it also a Label Distribution Protocol (LDP, RFC 3036) is running (see Figure 4.12.2). The routing protocol establishes - in a way described in subchapter 4.11 - in all LRs similar routing tables as it did in non-MPLS routers; then recorded IP address prefixes shall be mapped to labels which will enable packet switching inside the MPLS cloud. The simplified description of the procedure is the following: the ingress LER sends an LDP Label Request message to its subsequent LR identified in the routing table for every IP address prefix recorded for this next hop (see Figure 4.12.4). The subsequent LR allocates a label on the given interface to the address prefix in question. The label distribution may be independent, in which case the label is returned to the requester in a Label Mapping message directly. In case of ordered mode (see Figure 4.12.3) the LR waits with the answer until it receives a label for the same address prefix for its output interface (from the LR next to him). The egress LER will return the selected label at once. In this way, using the address prefix as a hook between the incoming label and interface and outgoing label and interface each LR sets up label forwarding information base (LBIF) for each IP address prefix (network address).

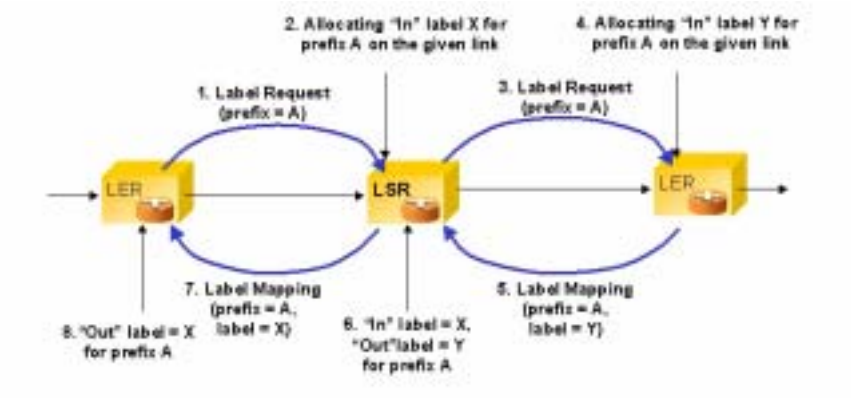


Figure 4.12.3 Label distribution in ordered mode

The forwarding function works as follows: the ingress LER classifies incoming IP packets (see DiffServ and IntServ) and determines the output interface and label for the packet. It attaches the label to the packet and sends it. On receiving the packet the LSR looks up the output interface and label in the LBIF, changes the label and forwards the packet to the next LR. The egress LER removes the label and sends the packet to the relevant outgoing (IP) interface taking into account the QoS requirements.

MPLS will function in a little different way when used for IP VPN. In this case the individual VPNs may use private IP addresses which may overlap between different VPNs and also data security is a more reflected function. Below we show one possible implementation of VPN using MPLS given in RFC 2547 but other solutions are not excluded (e.g. RFC 2917). The service provider allocates a unique destination VPN identifier to each VPN. Except this an 8-byte route distinguisher (RD) is allocated to each interface of the VPN, where it is connected to the public IP/MPLS network. Both destination VPN identifiers and route distinguisher shall be unique within an MPLS cloud even if it spans over multiple service providers. Within the MPLS cloud the routing protocol is advertising VPN-IP addresses instead of subnet address prefixes as address prefixes of different VPNs may overlap. The VPN-IP address is composed from the route distinguisher and the address prefix itself. Basically same route distinguisher could be allocated to each interface of a single VPN as normally subnet address prefixes within an Intranet does not overlap. However the MPLS cloud providing the VPN service and the Intranet using it belong to different administrative domains and the use of distinct RDs per interface defends the service provider's network from possible addressing errors in Intranets. The route

distinguisher also can be used to identify separate routes to the same subnet. As shown on Figure 4.12.4, each LER maintains a separate routing table for each VPN interface, which may be a physical interface or a virtual one like an ATM virtual channel, containing static dynamic records of the address prefix(es) used on the subnet connected, and allocates a VPN label (different from the MPLS label!) to it. In the static case address prefixes are entered by the VPN service provider according to the information received from the subscriber, in the dynamic case they are discovered by the LER itself using the routing protocol employed on the subnet. IGPs support a single table only and does not transport special VPN identifiers, hence are unsuitable for VPN routing. For this reason within the VPN-MPLS cloud an EGP routing protocol shall be used like the multiprotocol BGP-4 (see subchapter 4.11). Using BGP-4 each LER advertises its own IP address and all registered VPN-IP addresses together with corresponding target VPN identifiers and VPN labels. Other LERs also serving a VPN with the same target VPN identifier records these data into the routing table of the VPN; those not serving the given VPN just ignores the information. In the next step LERs request labels for the IP addresses of another LERs (and not for the address prefix of the subnet!) serving the same VPNs as they do by means of LDP.

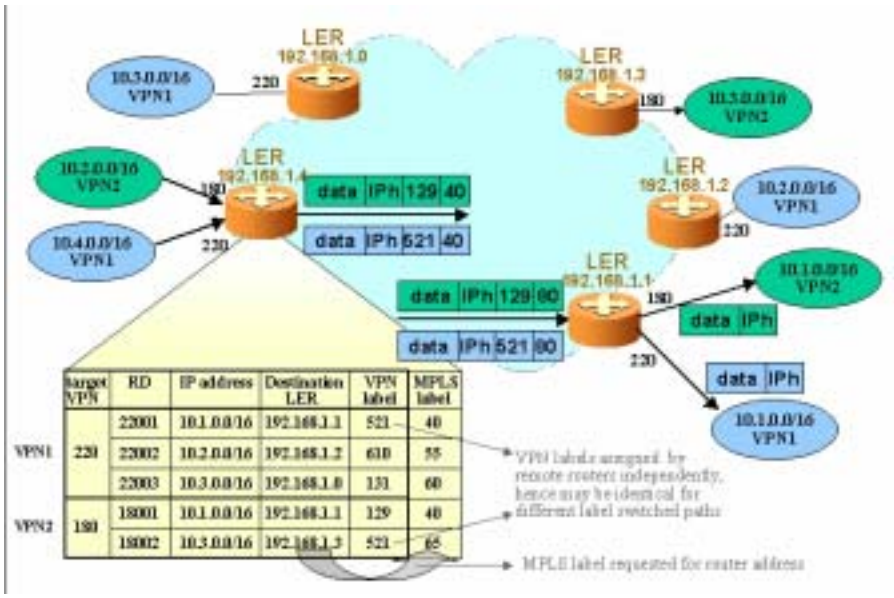


Figure 4.12.4 MPLS VPN routing

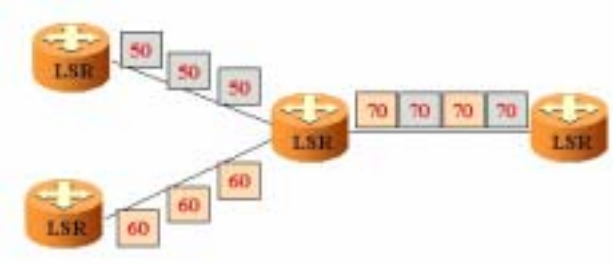


Figure 4.12.5 The ATM multiplexing problem

At data transfer the LER attaches two labels to the IP packet: the inner is the VPN label received from the remote LER for the destination subnet address prefix, the outer is the normal MPLS label associated with the IP address of the remote LER itself. Switched by the outer label the packet reaches the remote LER which, from the label associated to its own address, recognises that the packet belongs to a VPN. Using the inner VPN label the LER identifies which VPN the packet belongs to and after removing the VPN label forwards the packet to the relevant interface.

MPLS may be well combined with ATM, because functions can be split: routers perform control functions only and ATM switches handle the real traffic. IP QoS classes can be mapped to ATM traffic classes, LSR functions are executed by a combination of an ATM switch and a controlling IP router and LSPs become ATM virtual channels (VCs). In this scenario the IP-MPLS traffic will differ from other ATM

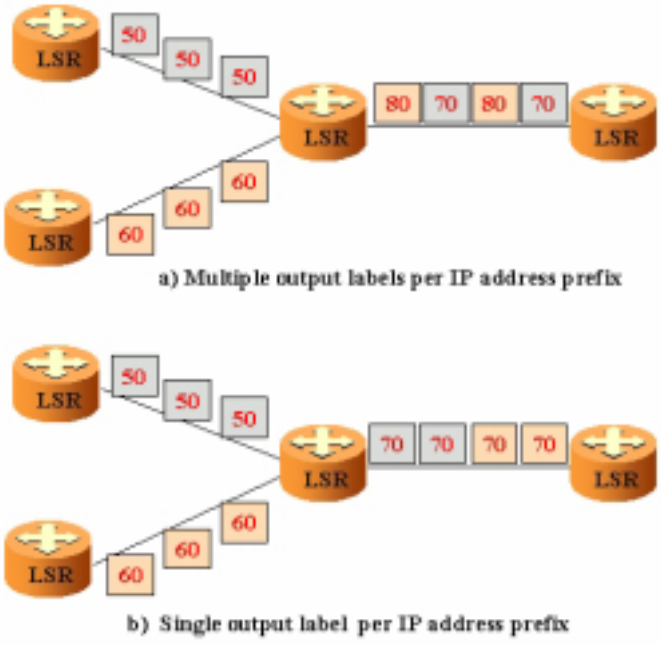


Figure 4.12.6 Solutions of the ATM multiplexing problem

1. For a given address prefix all LRs shall request different next labels for each incoming interface (see Figure 4.12.6 a)). As these labels identify different ATM VCs, no interleaving occurs.

2. The intermediate LSR gathers cell belonging to the same AAL5 PDU on input interfaces, then multiplexes AAL5 PDUs into a single outgoing ATM VC (i.e. uses a single label per address prefix, see Figure 4.12.6 b)).

4.12.6. Integration on the basis of 3rd generation mobile networks

Vast majority of users are connecting today to the Internet by a wireline solutions but forecasts for the near future predict significant increase in the number of users connected via mobile networks. Mobile Internet access technologies as GPRS (see subchapter 4.9) and 3rd generation mobile networks (3G) like UMTS (Universal Mobile Telecommunication System, one of the European members of the IMT-2000 family) provide sufficient bandwidth and expected to boom during the next a few years. However, for economical reasons UMTS will be able to spread step-by-step, in most cases migrating with current GSM/GPRS networks. The technology to enable this migration is the BICC.

The BICC (Bearer independent call control) architecture was specified in ITU-T Study Group 11 to allow shifting current PSTN/ISDN and GSM networks from the current TDM backbones to ATM/IP connectivity networks. It can not be forgotten that neither H.323 nor SIP/SDP support the portfolio of basic-, tele- and supplementary services of current PSTN/ISDN and GSM networks. On the other hand side BICC does not support future multimedia services. Therefore BICC is not an alternative for H.323 or SIP/SDP but supplements them by integrating “conventional” services with new VoIP and multimedia services over a common IP backbone network. Application of BICC in 3G networks is shown on Figure 4.12.7. The essence of BICC is the separation of the call control and the bearer control (signalling traffic) related to a single call. In this respect BICC is much like H.323. Existing PSTN/ISDN or GSM exchanges remain as controllers only (also called a telephony server) providing signalling support for call set-up, release and supplementary services but not switching user channels related to these calls (named the bearer). BICC exchanges also controls media gateways (MGWs) using the MEGACO/H.248 protocol, which are responsible to establish and release user traffic channels/flows and are ATM

switches and/or IP routers (BICC enables of fully mesh ATM and IP MGWs). Any suitable protocol can be used to control the bearer like the ATMF UNI, B-ISUP, SIP/SDP, H.323 etc. There is no restriction on this respect in BICC, the only limitation is, that it shall be capable of carrying the binding information tying together the call and the related bearer at both ends. The most important part of BICC is on the call control plane, where the BICC protocol is used. The BICC protocol is an adopted ISUP used today in NNIs (Network Node Interface) of PSTN/ISDN and GSM networks. Vast majority of added BICC information is carried within a transparent for ordinary ISUP container minimising modifications needed to existing exchange software.

The real future of mobile IP communication lies in 3G networks. At the time of writing this book the first commercial UMTS network (with restricted number of users) is put into service in Japan and several pilot is going on successfully in Europe. The brand new component of today's UMTS network is UTRAN (UMTS Terrestrial Radio Access Network). The most important feature from the point of view of this section of this new, ATM based 1900 MHz radio access network (see also subchapter 4.9) is the support of conventional (e.g. dedicated to the mobile user) and multiple access packet radio channels. Therefore it enables to access current GSM and ISDN services as well as direct IP communication at the speed up to 2 Mbit/s, what is a significantly higher rate then allowed by today's public radio access networks. The

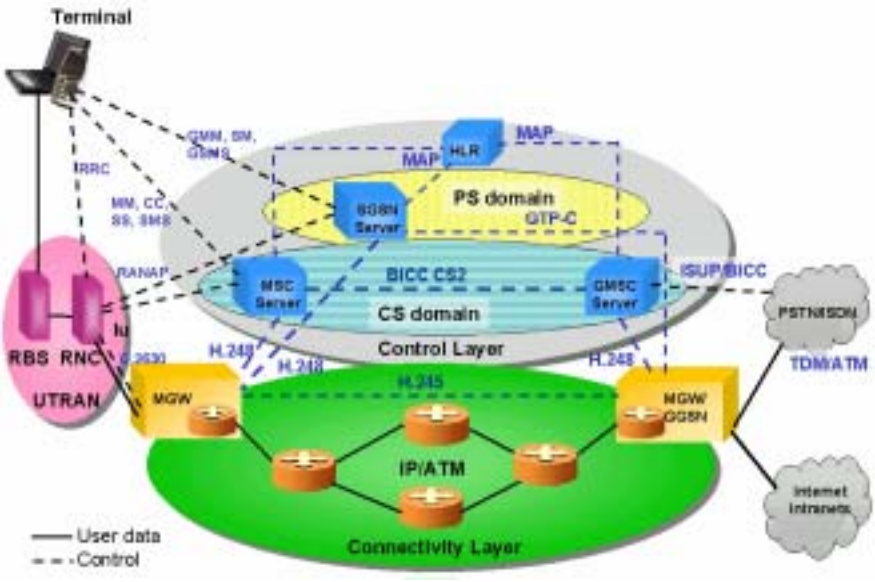


Figure 4.12.7 Using BICC in 3G mobile networks

current UMTS core network is composed of two parts: the circuit switched (CS) domain and the packet switched (PS) domain, which are basically identical to the current GSM and GPRS core networks. Both are completed with an ATM interface and signalling protocols needed to attach UTRAN. Important to mention, that the PS domain supports IP data traffic only, no real time traffic in the current phase.

As shown on Figure 4.12.7, the control plane servers and the MGW are connected to UTRAN on different interfaces. While MSC and SGSN (see also subchapter 4.9) communicating with UTRAN with the standard 3G signalling protocol RANAP (Radio Access Network Application Part) via the interface used today, MGWs are connecting with ATM trunks and the AAL2 signalling protocol (Q.2630) needed to set-up and release AAL2 paths. Existing mobile nodes, except handling current GSM/GPRS traffic, also provides signalling support for 3G traffic, while 3G speech and CS-data traffic uses an ATM/IP transport network integrated with the IP traffic. The service portfolio granted is fully corresponds to usual GSM/ISDN and GPRS capabilities. New protocols to be implemented are BICC and H.248, except them MSC and SGSN uses conventional L3 access protocols in a slightly extended for 3G form. A more detailed description of BICC CS1 (capability set 1) is given in [4.12.42], [4.12.43] also provides information about CS2 capabilities.

It is believed that in the course of time IP traffic will be prevailing also in 3G networks. For this reason research, standardisation and development of all-IP 3G systems has been started. One of the possible architectures of it is given on Figure 4.12.8. Similarity to the BICC architecture is remarkable although in spite of architectural correspondence, there are important dissimilarities. All-IP 3G networks will directly support IP multimedia communication which is not supported by BICC and current UMTS call control protocols. This will require significant technical development in many areas. One of the most important is the use of real-time routers, developed to be suitable for special 3G requirements (e.g. the proportion of the small packet size real time traffic is 85-90 % at the beginning) in UTRAN and in the connectivity plain of the CN. Desirable performance can only be achieved by combining different traffic management solutions (e.g. DiffServ to handle traffic of different priorities, fragmentation of long packets to decrease voice delay, i.e. RFC 1990, RFC 2686) and by managing network resources (e.g. bandwidth brokering, see subchapter 3.7). Also additional functions like transparency of STM/PDH

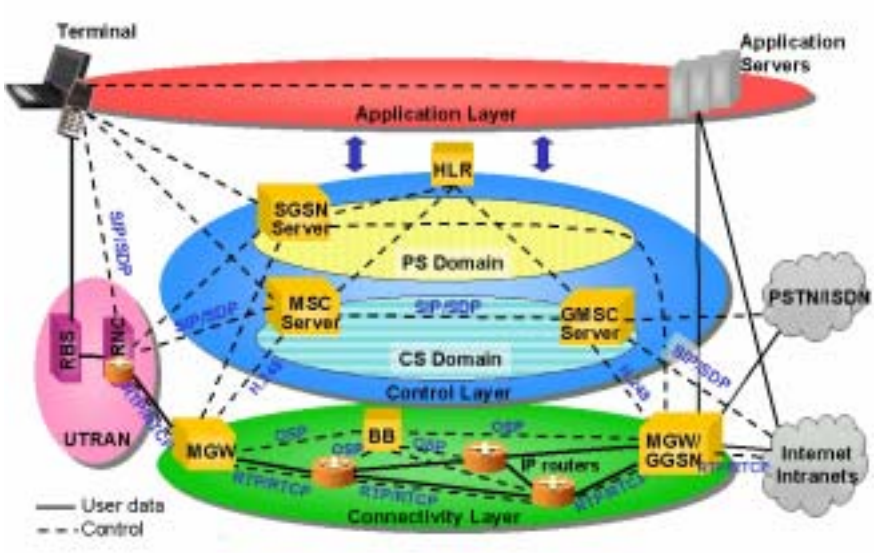


Figure 4.12.8 Network architecture of an all IP 3G mobile network

synchronisation signals shall be fulfilled. A standard call level signalling protocol shall be introduced to support multimedia calls – 3GPP has chosen SIP/SDP for this purpose – and the efficient transfer of IP packets on the restricted-bandwidth and high-loss radio channels shall be solved (e.g. with robust header compression). The multimedia subsystem of 3G networks will support IP v6 only (current PS domain supports in theory both IP v4 and IP v6). Therefore gateways will be needed to interwork with IP v4 networks (by header mapping) or to handle the communication of mobile IP v6 hosts via IP v4 networks (by tunnelling, see subchapter 4.11).

It is expected, that 3G IP multimedia services will appear within a few years, but also very likely that it remains a subsystem for a long time within 3G networks. In other words the GSM 900 and 1800 MHz speech and circuit mode data traffic and the non-IP traffic of 3G systems will represent a significant portion that necessitates the coexistence of traditional, BICC and all IP solution for a long time.

References (to subchapters 4.11 and 4.12)

[4.12.1] Martin W. Murhammer, Orcun Atakan, et.al., "*TCP/IP Tutorial and Technical Overview",* IBM International Technical Support Organization, 1998.

[4.12.2] William Stallings, Paul Mockapetris, Sue McLeod at all, "*Handbook of Computer Communications Standards; The TCP/IP Protocol Suite, Volume 3*", Second Edition; Howards W.Sams &Company, 1989.

[4.12.3] RFC 791 *"Internet Protocol"*, J. Postel, September 1981.

[4.12.4] RFC 792 *"Internet Control Message Protocol"*,J. Postel, September 1981.

[4.12.5] RFC 793 *"Transmission Control Protocol"*, September 1981.

[4.12.6] RFC 768 *"User Datagram Protocol"*,J. Postel, August 1980.

[4.12.7] RFC 1058 *"Routing Information Protocol",* C.L. Hedrick, June 1988.

[4.12.8] RFC 1241 *"Scheme for an internet encapsulation protocol: Version 1",*R.A.Woodburn, D.L. Mills, July 1991.

[4.12.9] RFC 1633 *"Integrated Services in the Internet Architecture, an Overview"*,R.Braden, D. Clark, S. Shenker, June 1994.

[4.12.10]RFC 1700 *"Assigned Numbers",* J.Reynolds, J. Postel, October 1994.

[4.12.11]RFC 1723 *"RIP Version 2 - Carrying Additional Information",* G. Malkin, November 1994.

[4.12.12]RFC 1727 *"A Vision of an Integrated Internet Information Service",* C.Weider, P. Deutsch., December 1994.

[4.12.13]RFC 1771 *"A Border Gateway Protocol 4 (BGP-4)*, Y. Rekhter, T. Li., March1995.

[4.12.14]RFC 1889 *"RTP, A Transport Protocol for Real-Time Applications"*,H. Schulzrinne, S. Casner, R. Frederick, V.Jacobson, January 1996

[4.12.15]RFC 2002 *"IP Mobility Support",* C. Perkins., October 1996

[4.12.16]RFC 2003 *"IP Encapsulation within IP*", C. Perkins., October 1996

[4.12.17]RFC 2004 *"Minimal Encapsulation within IP"*, C. Perkins: October 1996.

[4.12.18]RFC 2205 *"Resource ReSerVation Protocol (RSVP) "Version 1 Functional Specification",*R. Braden, Ed., L. Zhang, S. Berson, et all, September 1997

[4.12.19]RFC 2210 *"The Use of RSVP with IETF Integrated Services",*J. Wroclawski, September 1997

[4.12.20]RFC 2211 *"Specification of the Controlled-Load Network Element Service",*J.Wroclawski., September 1997.

[4.12.21]RFC 2212 *"Specification of Guaranteed Quality of Service",* S. Shenker, C.Partridge, R. Guerin., September 1997.

[4.12.22]RFC 2327 *"SDP, Session Description Protocol"*, M. Handley, V. Jacobson, April 1998.

[4.12.23]RFC 2328 *"OSPF Version 2"*,J. Moy, April 1998.

[4.12.24]RFC 2330 *"Framework for IP Performance Metrics",* V. Paxson, G. Almes, J.Mahdavi, M. Mathis, May 1998.

[4.12.25]RFC 2400 *"Internet Official Protocol Standards"*, J. Postel, J. Reynolds, September 1998.

[4.12.26]RFC 2474 *"Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers"*,K. Nichols, S. Blake, F. Baker, D. Black, December 1998.

[4.12.27]RFC 2475 *"An Architecture for Differentiated Service",*S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, W. Weiss, December 1998.

[4.12.28]RFC 2543 *"SIP, Session Initiation Protocol"*, M. Handley, H. Schulzrinne, E., Schooler, J. Rosenberg, March 1999.

[4.12.29]RFC 2547 *"BGP/MPLS VPNs"*, E. Rosen, Y. Rekhter, March 1999.

[4.12.30]RFC 2597 "*Assured Forwarding PHB Group"*, J. Heinanen, F. Baker, W. Weiss, J. Wroclawski, June 1999.

[4.12.31]RFC 2598 *"An Expedited Forwarding PHB",* V. Jacobson, K. Nichols, K. Poduri, June 1999.

[4.12.32]RFC 2893 *"Transition Mechanisms for IPv6 Hosts and Routers"*, R. Gilligan, E.Nordmark, August 2000.

[4.12.33]RFC 2917 *"A Core MPLS IP VPN Architecture",* K. Muthukrishnan, A. Malis, September 2000.

[4.12.34]RFC 2998 *"A Framework for Integrated Services Operation over Diffserv Networks"*, Y. Bernet, P. Ford, R. Yavatkar, at all, November 2000.

[4.12.35]RFC 3031 *"Multiprotocol Label Switching Architecture",*E. Rosen, A.Viswanathan, R. Callon, January 2001.

[4.12.36]RFC 3035 *"MPLS using LDP and ATM VC Switching",*B. Davie, J. Lawrence, K.McCloghrie et all., January 2001.

[4.12.37]RFC 3036 *"LDP Specification",*L. Andersson, P. Doolan, N. Feldman et all., January 2001.

[4.12.38]ETSI TS 101 329-2 v2.1.2 (2001-07), "*Telecommunications and Internet Protocol; Harmonization Over Networks (TIPHON); End to End Quality of Service in TIPHON Systems; Part 2, Definition of Speech Quality os Service (QoS) Classes"*

[4.12.39]ITU-T Recommendation H.323 (11-2000), "*Packet-Based Multimedia Communications Systems"*

[4.12.40]ITU-T Recommendation H.225.0 (11-2000), *"Call signalling protocols and media stream packetization for packet-based multimedia communication systems"*

[4.12.41]ITU-T Recommendation H.245 (11-2000), *"Control protocol for multimedia communication"*

[4.12.42]M. Oskar van Deventer, Iko Keesmaat, Pieter Veenstra, *"The ITU-T BICC protocol, the vital step towards an integrated voice-data multiservice platform"*, IEEE Communications Magazine, Vol.39, No.5, pp.140-145, May 2001.

[4.12.43]Réthy György, "*BICC, hálózatintegrálás szélessávon"*, Magyar Távközlés, 2001, 1.szám

[4.12.44]3GPP 23.821 v1.0.1 (12-2000), "*Architecture Principles for Relase 2000"*

[4.12.45]3GPP TS 23.922 v1.0.0 (12-2000), *"Architecture for an All IP network"*

[4.12.46]3GPP TS 24.228 v0.1.0 (12-2000), "*Signalling flows for the IP multimedia call control based on SIP and SDP; stage 3"*

[4.12.47]3GPP TS 24.229 v0.0.6 (12-2000), "*IP Multimedia Call Control Protocol based on SIP and SDP; stage 3"*

[4.12.48]ITU-T Recommendation G.109 (1999): "Definition of categories of speech transmission quality".

5. Services

Remarks of editor

The chapter 5 has a particular role in the structure of this wide scope book.

The networks have traditionally a technical oriented discussion mode because networks are very sophisticated systems utilising the latest results of technology sciences. Applications also have important role by giving reason to use networks and services. The issues of network operation an service provision based on the general telecom regulation system also have extreme importance to establish the general rules on the telecom market including technical and competition aspects.

Service provision is particular because this is the activity forming basis to send bills to the customers. The incomes from services are shared among the members of the whole activity chain including network operators, service resellers, system suppliers, construction companies, system developers, investors, researchers, regulation staff, etc. Dear reader! Please do not hope an economic discussion in this chapter! Detailed market analysis or price calculations are not involved in this part. Authors of this chapter are engineers and the target readers are engineers as well.

The services are discussed in the first part focusing on the roles in the activity chain, The second part systematically describes services as there were designed based on service capabilities of networks. A separate part of this chapter discusses a big sort of other services as products of innovative operation of networks beyond the original intention of system designers. The daily network operation can produce even richer and more beneficial services overperform the goals of technology-oriented researchers and developers. The questions of information content provision are not in the scope of this book. The access issues of content oriented services have so high importance so one part of this chapter discusses the main questions.

The network services are emphasised in a separate part because aspects of network interconnection, service provision to an other network operator or to big users having their own private networks form a specific set of service issues.

Terminal equipment are inherent part of networks but we discuss some terminal issues here because implementation of the higher and higher portion of

services is involved in the very intelligent terminal equipment due to the technology development.

The last part of this chapter discusses aspects of service and application interfaces. Application Programming Interfaces offer efficient and standardised solutions to harmonise services and applications.

György Takács PhD, Editor of the Chapter

5.1. SERVICE PROVIDER AND USER ROLES

Sándor Szilágyi dr., author

György Takács PhD, reviewer

In case of general treatment of the services - from both statistical and regulatory point of view - it is important to determine the activity area of the service provision; who the service provider is and whom it provides the service. By *service* we usually mean that somebody or a company is being paid for its activity carried out in favour of somebody else without any change in the ownership of facilities concerned. While examining the definition in general, we analyse the telecommunications activity itself together with its attributes and conditions, then its form in which the activity is done as a service provision.

Later on, we will examine the participants of the service activity who form three groups: those who provide the service, i.e. service providers (SPs), those whom the service is provided and those who are the beneficiaries of the service. The third group does not always correspond to those whom the service is provided.

5.1.1. Telecommunications activity

The word *telecommunication* itself does express some kind of activity. According to the usual formulation, telecommunication means message or data transmission in form of electromagnetic signals (hereafter we consider data as message, i.e. "datamessage"), which may include text, drawing, static or motion picture, speech, music, data bulk, computer program etc..

The first question to raise is the distance of the transmission: what distance can be considered to be a "real" *tele*communication? Is the signal transfer happening inside a computer or between computers or a computer and its peripherals, a kind of telecommunication? It seems to be evident that a long distance telephone call is made by means of the telecommunication activity but if at a meeting, the signals of a microphone are transmitted to the amplifier and further on, to the loudspeakers may be categorised as telecommunication? What is the meaning of the prefix "tele-"?

This is no means an academic question since the enormous effect of telecommunications activity on the society makes it necessary to define rules and legal obligations. It can be declared that there is no given distance beyond which the activity would be defined as telecommunication and below which it could be qualified as a simple signal or data management. E.g.: if two persons working at the same company are talking in two neighbouring rooms through a telephone exchange, it is telecommunication, but receiving some radio signals from a far nebula trillions of kilometres away in the space, is usually not considered to be a telecommunication.

Telecommunication should be differentiated from other usage of electromagnetic signals. In the telecommunication the signal is put on a telecommunications network (in a simple case, on a telecommunications circuit) in a transformed appearance compared with their original form due to the nature of transmission. The primary aim of this transformation is to retain the original content of transmitted message while the transmission medium cause the minimum distortion, noise, bit error or other deterioration.

Thus, the transmission of signals by means of radio waves is always a telecommunications activity, happening either between the Earth and a spaceship or a driver and the central lock equipment of his car, parking 1-2 meters away from him. It can be observed that the radiocommunications network makes such transformation always necessary; namely radio waves are processed by modulation from electric signal current or voltage appearing in the metallic conductor which signals convey information to be transmitted.

Telecommunication is the transmission of speech information acoustic signals in electromagnetic form through the subscriber's line where. among other factors, their symmetrical transmission, their amplification by the Central Battery power, the matching between the telephone xchange and telephone set, all serve for bridging the distance between the subscriber and the telephone exchange with the minimum of deterioration. A typical implementation of electroacoustic system when the amplifier sends signals with increased voltage to the loudspeakers in order to decrease the transmission losses (so called 100-Volts system) can also be called telecommunication.

Naturally, beside this essential element there are other characteristics of telecommunications activity. The direction of signals also plays an important role in

certain forms of telecommunications. We can talk of addressed and unaddressed telecommunications; an example for the first can be the telephone service or the Internet while another example for the last one is the (uncoded) radio broadcasting. The whole procedure of addressing, including the reception, handling and usage of the address for routing the message towards the destination, is another part of the telecommunications activity. Similarly, establishing a virtual or real connection, its maintaining, modification or clearing, likewise - if necessary - settling the tariffs, also contribute to this activity. Temporary storage, coding of the signals with the content unchanged are belonging to the same area.

Here we have to mention some important definitions related to telecommunications as an activity. Their detailed discussion can be found in other chapters; here we are confined to a brief summary only. One of them is the telecommunications network which incorporates a set of transmission, switching, repeater, router, storage, gateway etc. facilities which serve for transmission and/or reception of messages to and from any network termination point in accordance with pre-determined, unified rules within the physical space or territory limited by their capabilities. The devices can be both hardware and software elements and the former in almost all practical cases are programme controlled. In addition, there are scarce resources belonging to the network such as frequency bands used by the network as well as finite stacks of addresses i.e. names. In the opinion of many experts, the public territory occupied by the telecommunications paths belong to those scarce resources too.

There is another important definition: telecommunications infrastructure. Besides the network itself, it includes the terminal equipment used by the customers and human and machinery resources of maintenance, repair as well as the telecommunications buildings.

It is a commonplace, that convergence can be observed in the field of the telecommunications. In relation of the network usage, earlier the public and the private networks were sharply separated; however, today those are merging in numerous cases so that the same network can serve public, business or private telecommunications activity at the same time. On the other hand, the borders between the previously defined sectors of the telecommunications such as telephony, data transmission and broadcasting are fading away and the primary sectorial

purpose of the network does not exclude other kinds of telecommunications activity. We discuss it in the next chapter.

Convergence can be also observed in the area where the telecommunications activity is tightly interwoven with the information technology. Databases of very large volume are built in the networks, computers with sophisticated operation systems are connected to the switching equipment for enhancing the network management and the maintenance.

5.1.2. Telecommunications services

The telecommunications activity appears in the form of services when the activity is performed for other people in return of a fee. The telecommunications activity appears rarely as a standalone service. In most cases the telecommunications service is a common result of several activities. An important question whether the service could appear as a standalone service when the telecommunications were extracted? We consider those services as telecommunications ones which can be used only together with or by using the telecommunications activity. Thus the telecommunications service has an including concept.

Here we must stop for a while because in the Hungarian language a common word is used for the *service* and the *service provision*. On the field of frequency management which is aimed at the activity instead of the service provision, a long time used word for the utilisation of radio frequencies for different purposes - the *radio service*. In the legal framework of the telecommunications this Hungarian word is used only in the formulation of *emergency services*. However, in the context of technology primarily in the field of standardization, the interpreters are frequently using the same word for description of technical capabilities of a certain infrastructure in the text of a standard. For example, The standards of Integrated Services Digital Network (ISDN) describe the technical capabilities i.e. requirements of that network while the Code of Practice of telecommunications companies is talking of Integrated *Service provisioning* Digital Network.

Further, in the Hungarian language those concepts expressed in English by words *service, facility, feature* are also attributed by the single word *service*. (English

language documents are not consequent in usage of those words, too. Recently in the European *acquis communautaire* the use of expression *service provision* is spreading.)

Therefore, it is reasonable to differentiate the *basic services* and the *supplementary services* relying upon the former ones. The basic services are a set of services which, when omitting any element of it, the minimum criteria included in the relevant decree or it does not comply any more with the standardised or broadly accepted definition of the service. The network is capable to provide this service together with the simplest type of terminal equipment attached. In the case of connection-oriented telecommunications, usually the origination of establishing the connection path, its maintaining, clearing, modification; in the case of connectionless telecommunications, the acceptance, routing and termination of messages belong to the category of basic services, together with the charging functions related to those activities.

The supplementary services, called often as additional services, as a rule, enhance the value of services provided for the users, therefore the service providers are collecting some additional fee for their usage. It is accepted to define them as *value-added services*. Characteristic examples of those are the telephone calls with number and/or tariff translations such as the "blue number" (numbers to be called nation-wide with local tariff), the "green number" (the call is paid by the called party), the call diversion etc.. Certain informatics and/or content services are also called value added services which enable the caller to get information, entertainment, transaction opportunities (e.g. remote polling, usage of banking, commercial services etc.).

In general, the basic services can be categorised, as a rule, in two classes: access services and traffic services. The access service is that part of the telecommunications services which provides the user, as a matter of principle, an opportunity for the usage of the network. The traffic service is the real usage of the network for conveyance of messages. The fees charged by the service provider reflect this duality: the installation and subscription fee is paid for the access while the connection and per-minute fee is charged for the traffic.

There is a case when the user pays only access fee: such case is the leased line service related to a particular circuit as well as the flat-rate services (e.g. cable

TV subscription, Internet access without any restriction related to data volume). Virtually, only traffic access fee is to be paid for the use of a public payphone, however, a part of the fee paid covers the expenses of installation and the maintenance of operability and on this account the per-minute charge is higher than that for the residence telephones.

As it is mentioned above, there is a decisive issue whether the service provided through the telecommunications network and including elements of informatics and content can be utilised without the use of telecommunications network? An indispensable part of radio broadcasting is the real-time programme transmission; the news or the weather forecast in the form of book or compact disk make sense likely as a historical record. However, in another case, a commercial purchase to the debit of the telephone card does not require the telecommunications network as a prerequisite since that transaction may also take place by the use of a normal credit card or cash.

It is worthy to investigate from this aspect the Internet where the role of telecommunications is relatively small. Even the Internet service providers are using such an expression that they are access providers. This opinion is reflected by the fact that the Internet, which is called popularly but not too properly "network of networks", performs only the maintaining its integrity and conveying it to the destination, however, the well-known Internet services such as surfing, electronic mail, file download, chatting are the merits of the servers which are connected, as a matter of fact, as terminals to the network. At the same time, the services of those servers cannot be utilised without the Internet network compiled of local area computer networks (LANs) and routers. Here we meet a case when the basic telecommunications service is scarcely used and the value-added services are dominating.

Another way of classification of services is to divide them to *bearer services* and *teleservices*. More detailed description is to be found in other chapters.

A further manner of classification is based on whether the transmission is realised in real-time or with intermediate storage. The service considered real-time means that the information is conveyed *in statu nascendi*, in the moment of its birth or with the original time relations of the event to the user. Here a storage for a short time interval occurs, of course, too (e.g. in the time-division switching stage of the

modern digital telephone exchanges) but it is negligible enough for the user could feel to be virtually on the site, i.e. the changes conveyed by the transmitted signals seem to be natural occurrences. Those classes are shown - in a very comprehensive, technology-dependent categorisation - in the following table.

Class of telecommunication	Real-time	With intermediate storage
Addressed duplex	Speech (telephony, mobile radiotelephony), teleconference, voice and data bearer services, interactive audiotext	X.25, facsimile, Internet surfing, chatting, e-mail, file download
Addressed simplex	Subscribed and/or coded broadcasting, news gathering (SNG)	SMS, paging, propriety protection systems, voice mail services
Unaddressed	Non-coded broadcasting, SRD devices	-

It should be noted, that there is a definition of the telecommunications services which excludes the broadcasting service [5.1.1]. This in unacceptable due to the fact that the broadcasting is a telecommunications activity and Hungarian statutes have considered them for the past decades in a similar way.

5.1.3. The service provider

The service provider may be au undertaking, but not necessarily is an operator or an owner at the same time. In the everyday economic life these roles are often separated. In many developed countries one makes a difference between the network operator and the service provider and a separate licence is issued for the one or the other. The latter one (SP) is handled as an undertaking which enters the marketplace without own infrastructure and the expression sometimes receives a pejorative stress. On the other hand, the word "public" has a peculiar history; in the beginning it had been used for characterising not the group of users but for the ownership of the network ("state-owned network") and this fact retained the Hungarian legislators from using this word for the networks.

The service provider can be a natural person or a legal personality or an organisation without legal personality providing its services on the basis of its fundamental act and the registration of its firm. A precondition of its operation is to have in a checkable way those material and human resources which are necessary for the provision of services. The material resources include its own or hired network, with the licences for the telecommunications lines running on anothers territories, the

licences of radio stations operated or hired (which is based on valid frequency assignment), the usage permission of the telecommunications buildings as well as the right of use of scarce resources needed for the operation of its own network or service.

The service providers may carry on their activity upon notification, licence or concession contract. Details are to be found in the Chapter 8 on the regulation.

The designation and the content of each service should have an unambiguous linkage in order to exact understanding between the service provider, the regulator and the user. This is contributed by the code numbers of the List of Products (SZJ) of the Hungarian Central Statistical Office (KSH). The code numbers for the telecommunications services are based upon those of the European regulation [5.1.2] while it should be noted that the assortment of the codes is rather narrow and the designations are too general.

5.1.4. The users of the services

An important attribute of the telecommunications services is the subject for whom the service is provided. The expression "for whom" has two points:

- who pays for it?
- who is the beneficiary?

The objective of the telecommunications regulation is primarily the service provided for a broad circle of the society, to the benefit of the public. This kind of service is called publicly available service or shortly, public service. (The Telecommunications Act No. LXXII of 1992 had stipulated this kind of service as a "service for the purpose of the community" and the "public" word had been reserved for the payphones only. The Communications Act No. XL of 2001 does not use neither the "for the purpose of the community" nor the "public" expression.)

The public service can be characterised by the feature that the service provider is obliged to enter a so called subscriber's contract with his regular user, that is, the subscriber pays for the service. There are exceptions because some public services can be used without contract (e.g. use of a payphone, use of Internet in an Internet café etc.). Considering that the subscriber's contract is composed, according to a statute [5.1.3], of two parts: of the general conditions of undertaking (code of

practice) of the service provider and of the individual subscriber's contract; in the cases mentioned above only the written individual contract is missing but the user by means of the usage has recognised the general conditions of undertaking as binding for himself.

In principle, the public service can be used by anybody: private person or any kind of associations. If the service provider does not advertise his service publicly, does not publish its conditions, but he defines the circle of customers for which he wishes to provide the services, in reward of a fee, so this is a non-public service.

One of the public services is the telephone service when anybody can subscribe it, but it is a non-public service, when the telecommunications activity is carried out only for the workers of a certain company by operation of the company private branch exchange.

In some cases the service appears in a "public habit" but, having investigated who pays for the service, it can turn out that appearances are deceptive. One example is the broadcasting which is often classified improperly as a public service since the radio programme, the civil service TV programmes can be received by anybody within the area of reception. However, the situation is not so simple. The TV set operational fee which is collected by the state as a tax from the users of the TV sets has the purpose of supporting the programme production which is, in turn, not a telecommunications service. In the chain of *programme production - broadcasting - receiving* the programme producer pays for the broadcasting as for a telecommunications activity, consequently here a non-public service takes place. If anybody applies for broadcasting his programme, in vain, nobody would accept it because it belongs to the competence of the National Radio and TV Body.

Another, similar example is the interconnection service which takes place at the interoperation of different networks. In spite of the fact that in the case of telephony, the beneficiary of this service is the telephone subscriber, his local operator pays for the interconnection service. Although the statute on revenue sharing defines a certain part of the charges collected from the subscriber, which should be transferred to the interconnection service provider, however, this money flows to the local telephone service provider in the framework of the legal relationship of the subscriber in exchange for the national and international long distance calls.

Legal relationship is established only between the service providers, thus this service is non-public. In this category falls as well the call termination service.

The previous legal regulation has made a difference between the users of the public telecommunications services whether the user is a private person or an undertaking. Even the telephone tariffs have matched this difference; based upon socio-political considerations the residential subscribers paid lower installation and subscription fees than the business subscribers. However, this difference is essentially disappeared in 2002.

In certain cases the fee for the service is paid not by the real beneficiary of the service. One of characteristic examples is the "green number" service, charges of which is paid by the called party. Another example is the paging service. When somebody sends a message to his friend's pager receiver, the price of this service is paid not, or not only by him, but the subscription fee is charged to the person who bears the receiver on his belt or in his portfolio. The latter person pays for it even in the case when the message does not serve his interest. The sender of the message is charged only with the access fee by his telephone operator.

There is a curious mixture of public and non-public services when a non-public service is provided via a network used for the provision of public services. Three designations are used for this case.

First is the Closed User Group (CUG) which means a user's group formed among the users of a GSM or ISDN network. Its members are calling each other with non-public shorter numbers, have no legal relationship with the public service provider and the charges are paid, instead of them, by the undertaking they are working at. They can originate calls toward the public network by individual authorisation, dialling a prefix.

The second is a version for the PSTN: the Centrex service (if the members of the group served by Centrex belong to the same local exchange) or the Wide Area Centrex (if they may belong to different exchanges).

The third designation is the Virtual Private Network (VPN) which is a generalisation of previous two concepts. For the connection of the users of VPN can be used many kinds of networks (fixed, mobile, PABX). The VPN service provider translates the interim numbers dialled by the members of the group into public

network addresses and conveys the calls in such a way toward their destination. At the same time, addresses used by calls arriving from the public network are included into a special numbering field reserved for the VPNs.

An opposite case occurs when a broadcasted sideband is utilised (RDS - Radio Data System) for a purpose quite different from the original service, e.g. for paging. In this case an infrastructure aimed for non-public service is used for a public one.

As is it obvious from above, that the continuously enlarging circle of the services includes also elements of informatics and content. This fact has an effect on the legal relations of the services, too. In the content transmitted by the telecommunications - especially in the case of radio and TV broadcasting - the proportion of the real information to reclame can turn the legal relation upside down. In some forms of telecommunications, e.g. cable TV services, the service provider has to pay for the intellectual rights of the programmes while in other instances the programme producer may pay for the transmission (as it is always the case even at the broadcasting of the civil service programmes).

Some years before a service provider experienced with charge-free telephone calls when mixing advertising text into the conversation in defined regular times. Calls arriving at his "green number" were diverted back into the public network via an "advertisement-mixer" device. This procedure changed not only the direction of cash-flow but also inserted a non-public element into the public service, because the reclame provider paid for this part of the service.

This case demonstrates the significance which certain services may have for the society. It is not a neutral thing for the society that the broadcasting stations are operated at a high level of quality, including the operation of "advertisement-mixers" ensuring free of charge telephone calls, the undisturbed and secure operation of the telecommunications networks serving for the Automatic Teller Machines (ATM) with growing usage rates, since the benefits of those services can be enjoyed by any citizen who can enforce his/her interest only in a very complicated way. The future regulation certainly will not differentiate so sharply the public and non-public services.

References

[5.1.1] Directive 97/33/EC of the European Parliament and of the Council of 30 June 1997 on interconnection in Telecommunications with regard to ensuring universal service and interoperability through application of the principles of Open Network Provision (ONP), OJ No L268, pp. 37-62 (3.10.1998)

[5.1.2] Council Regulation (EEC) No 3696/93 of 29 October 1993, as amended by Commission Regulation (EC) No 1232/98 of 17 June 1998 on the statistical classification of products by activity (CPA) in the European Economic Community

[5.1.3] Governmental decree No 243/1997. (XII.20.) on the subscriber contracts in the telecommunications

5.2. Dialog Based Services

Erika Madarász, author

Magdolna Nándorfi dr., reviewer

Dialog is one of the most common forms of communications. Telecommunications give capabilities to communicate persons having any distances in between them. The most common telecom services supporting dialog are:

- telephone services transmitting human speech
- video telephone services transmitting human speech and moving picture simultaneously.

Nowadays the spreading of video telephone services has low level but the traditional telephone service seems to be the most frequently used service. Methods and principles elaborated for telephone services can be applied for video telephone service also, because the call set-up and transmission functions are similar and the networks used are partly common. The details of video telephone services will not be discuss below.

In this section the general characteristics of telephone services, the network requirements, and the supplementary services will be discussed in details.

5.2.1. Telephone Service

The telephone service provides two-way, symmetric, real-time speech transmission on the connection of telecommunications network between two users. The connection has been set-up for the duration of telephone call and has been initiated by dialling the phone number.

A telephone service can be characterized from technical and regulatorial point of view. The applied technology is in the focus of technical discussions. So telephone services can be provided in wire line and mobile telecom networks as well, within analogue and digital transmission techniques and circuit or packet switching solutions are also applicable. The service-oriented characteristics of different telecom networks will be discussed in point 5.2.4.

The telephone service can be characterised from the aspect of regulation by the possibilities and functions of service provision, with the relation of service provider and user, and the conditions of using services, independently from the telecom networks and the applied technology.

One of the most important features of services the group of users who are able to reach the service. Regulation aspects differentiate:

- Public telephone services (introduced in point 5.1.4) give the possibility for anyone to access services.
- Private telephone services providing communications between the members of a company or an organisation.

The provision of private telephone service is based on individual contract between the service provider and user organisation. Regulation for this activity is not needed. The partners can make an agreement about the legal, commercial and technical details of services. The private telephone services have individual elements so they are not uniform.

The importance of public telephone services has reached very important level in the society so the regulation is needed. First of all, for the users have to be granted the possibility to access the services. Secondly, the fair competition for the service providers needs also specific regulations. The mandatory elements of telephone services for service providers and the minimal quality requirements need detailed specifications. The public telephone services are discussed in detail in point 5.2.2 below.

5.2.2. Public Telephone Service

Public telephone services can be provided using both wireline and mobile telecom networks. Concerning services the user can access to the network, make call set-up, communicate with the partner, and pay the bill.

The service providers have additional obligations beyond the basic activities based on legal regulations. These additional obligations are part of the services and will be discussed below. The users can select supplementary and value added services according to the offer of service providers.

So the component of public telephone service can be separated in the following subsets:

- Basic mandatory elements
- Additional obligations
- Optional elements

5.2.2.1. Basic Mandatory Elements

The following elements are needed for basic activity of service providers:

- Call set-up of national and international calls to any telephone subscriber or any other telecom services (e.g. directory service, voice mail service) or value added services (e.g. speaking clock, time-table, account balance) initiated by telephone number.
- Emergency calls with short numbers and special call handling terminated in the special call centers of the ambulance, fire brigade, and police including the European harmonized emergency call number 112.
- Customer care service for any kind of administrative contact between the service provider and the customers including personal, telephone-based, e-mail, mail, telefax, etc. contact.
- Fault reporting for registration of any customer complain concerning the services or networks.
- Directory services (information on name, directory numbers, address of subscriber) by phone or phone book (paper or CD form).
- Billing: accumulated value of fees for used services. This sum can be sent to the customer in bill form or can be subtracted from a pre-paid account. Bill contains the subscription fee, and fees of the initiated calls and supplementary and value added services.
- Network management including operation, maintenance, fault clearing and development of network.

5.2.2.2. Additional Obligations

The most important additional obligations for service providers based on legal documents are:

- Public phones: operation of public phones (coin-boxes or card phones) in places of public access to initiate calls by any user, usually in the PSTN/ISDN.
- Legal interception for the security services.
- Carrier selection: possibility for customers to select a service provider to be involved for access to the called subscriber or service in form of *Carrier pre-selection* (the subscriber selects the involved service provider in the contract

and no carrier selection prefix is used for calling) or *Call-by-call carrier selection* (the subscriber can select for each call the service provider to be involved in the given call - other than the pre-selected service provider - by dialling a carrier selection prefix).

- Number portability: possibility for customers to resign their subscription with a service provider and to contract another subscription with another service provider without changing their number, and without changing the nature of the service offered.

5.2.2.3. Optional Elements

Supplementary and value added services are provided by telecom companies to make more attractive the services and fulfil the needs of customers. The subsets below select the services based on the character of services and the method of implementation:

- Subscriber supplementary services concerning the implementation and provision of services which are part of subscriber contract or other individual contract (see 5.2.3. below).
- IN services: usually special numbering schemes for special tariff schemes, pre-paid services, personal numbers, etc. The IN services are implemented with implementation of centralised intelligence integrated in the network (see 5.2.4.2 and 5.2.7 below). The service provider of IN services might be different from the normal telephone service provider.
- Value added services: beyond the telecommunication these services provide other services, e.g. information, bank transactions, access to services of other networks (e.g. Internet, paging, X.25). The telephone service provider gives access to the value added services.

5.2.3. Telephone Subscriber Services

The customer oriented definitions of subscriber services are the offered possibilities, capabilities, functions or operations provided by the telecom companies.

The public telephone services described in point 5.2.2 above can be used by the customers on the bases of customer contract except the usage of public phones. The general conditions of the subscriber contract contain the subscriber services, the mode of usage and the fees.

The subscriber might be a natural person or legal entity or enterprise without legal personality in a contractual relationship with a service provider. The subscriber pays the bill.

The user is the person who uses the telephone service and might be different from the telephone subscriber.

One part of the services is included in the subscriber contract so the subscriber can use them without extra procedure. The supplementary services have to be ordered from the service provider.

Basic service

The set of public telephone services including call set-up, emergency calls, directory services, phone book, customer care services, billing services, customer complains.

Subscriber supplementary services

The set of public telephone services provided on the request of the subscriber concerning call set-up, charging and operation and characterized by the following items:

- Supplementary services concerning call set-up are for the higher comfort of the users. The set of such services are highly determined by the switching systems and the capabilities of the networks.
- Supplementary services concerning charging might be special price packages or temporary actions depend mostly on the marketing strategy of the service provider. The technical aspects are not so crucial in this set of services
- The operational services like transfer the location of service, the change of subscriber on the same location, modification of contract cover mostly administrative task and less technical part.

The subscriber supplementary services concerning call set-up are discussed below.

The set of subscriber supplementary services may show high level of variety for different reasons. The set of services extended continuously by the development of technology and the market driven demands but their implementation in networks might be in different time-schedule and mode. A subscriber in a given moment is connected to the telecom network by a well-defined mode. The available set of supplementary services is influenced by:

- The network capabilities and the technology of the subscriber access (e.g. access to analog or digital network, or access to wireline or mobile network);
- The capabilities of the terminal equipments.

The telephone service can be implemented usually by the network (including the switching systems) and the terminal equipment together. Some services need special functions only in the network or only in the terminal equipment. The supplementary services based on terminal equipment functionalities like loud speaking operation, displays, or memories will be discussed in point 5.6. In this part will be discussed supplementary services based on network features and abilities.

5.2.4. Services Characteristics of Voice Oriented Networks

The call set-up and voice transmission functions are implemented by interconnection of global networks especially planned for voice transmission. Details of transmission networks can be found in Chapter 4. Here you can find only the summary of the characteristics that are relevant in the services.

The main forms for the voice oriented narrow-band circuit-switched networks are the PSTN/ISDN and the PLMN networks. These networks have unified numbering schemes without problems of interworking. The IN, which is connected to the PSTN/ISDN and PLMN networks, is able to introduce a big set of services based on centralised service control functionalities.

By the end of nineties the voice transmission on the packet-switched networks has fast rollout. The results of VoIP development have some feedback to the traditional telephone services as well.

5.2.4.1. Wireline Telephone Network

The PSTN - Public Switched Telephone Network and the ISDN – Integrated Services Digital Network are the main element of the global telephone services. Generally the PSTN was developed for the voice transmission but step-by-step its usage for data transmission increasing (telefax, modem based Internet access, etc.). The ISDN was generally developed for supporting different services based on ITU-T – International Telecommunication Union recommendations and ETSI – European Telecommunication Standardization Institute standards. However the telephone services seems to be the most common application of ISDN even in 2002. The ISDN standardised bearer and teleservices are discussed in point 5.3 below.

The PSTN and the ISDN networks are not separable because the ISDN switches and the digital interconnection of the switches transport the ISDN and PSTN traffic together. The main difference is the mode of subscriber access. The subscriber terminal equipment is different and supplementary services are different also. The PSTN supplementary services are listed in 5.2.5. and ISDN supplementary services in point 5.2.6.

The hunting group lines connecting PBX equipment are different to the individual line access concerning the services as well. The hunting group can be implemented using analog loops or ISDN access. The users of telephone sets connected to PBX lines can utilise only one part of the public switch telephone services and some cases with modified procedure. There are specific PBX features like direct-dialling-in. The PBXs provide special services also like abbreviated dialling, call transfer, call forwarding, call completion on busy, call pick-up etc.

The Centrex provides function like a PBX but is implemented in public network switches. The Centrex provides services for both analog and ISDN access and also mixed access if necessary. Concerning services the services the Centrex is similar to a PBX with direct-dialling-in. The users practically could not sense whether Centrex or PBX implements the functions.

5.2.4.2. Intelligent Networks

The IN – Intelligent Network is based on PSTN/ISDN and PLMN networks. One of its main principals is that controlling the intelligent network services is implemented in a separated element named SCP – Service Control Point. The database for the control functions is connected to the SCP. The switching functions are implemented in the ditital exchanges with and extended switching functionality named SSP – Service Switching Point. The voice path is always switched through the SSP switches of PSTN/ISDN or PLMN, but the call set-up is controlled by the SCP. Most common IN services are listed in point 5.2.7.

5.2.4.3. Public Land Mobile Networks

The most important characteristic feature of the PLMN systems is that the users can use the service from anywhere even when move on with their wireless terminal equipment where the network has coverage.

The mobility functions are supported by the following functions beyond the radio transmissions:

- Location: the ability of the terminal equipment to detect and decide by different measurements the appropriate radio cell where the terminal is.
- Roaming: the capability of the system to use services at locations where only a partner service provider has coverage
- Registration of position and individual data of terminal equipment and user in the system by the HLR (Home Location Register) and VLR (Visitor Location Register) to have full accessibility.
- Hand-over: maintaining the connection during the telephone conversation without interruption and quality degradation in cases when the user changes his position from one cell to other.

The number and importance of the first generation analog systems decreasing by 2002. In Europe the second generation GSM – Global System for Mobile Communication digital system has practically full geographic coverage and very high user penetration.

The GSM has been developed to support different services. ETSI standards describe very precisely the tele, bearer and supplementary services. The GSM supplementary services are very similar to ISDN supplementary services. The GSM supplementary services are listed in 5.2.8.

5.2.4.4. IP Networks

The Internet Protocol (IP) based packet switched networks were originally developed for data transmission. The results of the development make it suitable for voice transmission as well using the VoIP – Voice over Internet Protocol. The target of the developments is the integrated communications and multimedia communication and several organisations in the field of telecommunications standardisation has resources to solve the problems in the frame of IETF – Internet Engineering Task Force,

ETSI, ITU-T together with manufacturers.

Interworking of narrow-band circuit switched and IP based packet switched networks concerning telephone services has step-by-step development. The simplest scenario when the call originated and terminated in circuit switched network and IP based network is used only for transparent transmission of voice signal has widely used solution in the daily practice.

The full transparency of the two types of networks needs the voice signal transcoding but the most interesting task is the addressing scheme and transcoding of signals of call set-up.

The ITU-T has elaborated Recommendations for IP network based multimedia communication under the umbrella of Recommendation H.323. This Recommendation involves the call set-up and voice transmission between users having IP addresses and also communications between one user and the gateway having connections towards the PSTN/ISDN networks. The implementation of subscriber services (well-known in PSTN world) has been started in IP based environment and elaboration of interworking has been started as well.

The TIPHON project in the frame of ETSI has studied the interworking of PSTN and IP based networks concerning telephone services based on ITU-T Recommendations.

The IETF has elaborated a new protocol for call set-up in the IP world: SIP - Session Initiation Protocol. This protocol needs further development to support the full set of services.

The transparency of addressing procedure seems to be an important task, too. The ENUM is a protocol and domain name system making possible the access of IP networks users from PSTN network dialling an ordinary phone number. The ENUM has a protocol mapping the phone number into a domain name and look for the URL of the user having the given phone number. So the location and the mode of availabilty can be determined for the called subscriber.

The voice coding and the implementation of signalling protocols has now critical issues concerning telephone services in the cases exceeding a quality limit and do not touch the main features of the services. In the field of public services the subscribers connected to IP based networks needs practically the same set of basic

and supplementary services which are common in the traditional telephony: dialling by phone number from the IP network, accessibility of emergency services, possibility of legal interception, directory services, phone books, customer care and handling of customer complains (see 5.2.2).

5.2.5. PSTN Subscriber Supplementary Services

The supplementary services provided by the switches and the network for PSTN subscriber can be used with different telephone sets with different comfort level. The main enabling factor is the transmission of signals on the analog wires so we discuss first of all the different solutions in signal transmissions.

5.2.5.1. Signals on Analog Subscriber Loop

Traditionally the analog subscriber loop is a twisted pair of copper wires so this loop is suitable for transmission of voice band (300-3400 Hz), DC and out of band signalling.

DC signalling: Closing the loop initiated by off-hook of handset means the intention of user to start a call or other cases during the ringing signal means the intention of user to answer the incoming call. The on-hook the handset means the intention for ending the call. The pulse dialling means a special coding of telephone numbers with staggering the DC current. For supplementary services the Flash function, which generate a short interruption of loop DC current, is used.

Out of band signalling: Ringing the bell the telephone switch provide ringing current on the loop. The frequency of ringing current is usually 25 Hz. Different cadences can be used for supplementary services differentiate calls, e.g. call-back, held calls, and identification of call number at multiple subscriber number service. Out of band signal is used to indicate the start of charging in the form of 12 kHz pulses.

In-band signalling: DTMF – Dual Tone Multiple Frequency signals are used in the „tone dialling” of the telephone sets to send called digits to the switch or to provide additional information during the dialog with automatic systems. In-band signals are the feedback information for the users like dial tone, ringing tone, busy tone and speech announcements. In-band signalling protocol has been developed for display information on the telephone sets based on FSK Frequency Shift Keying

signal. The display can show the number of the calling subscriber, signal for indicating waiting messages in the Voice Mail System and even short messages like SMS in the GSM system.

5.2.5.2. Supplementary Services on Analogue Subscriber Loop

PSTN supplementary services has no detailed standards so different suppliers offer different set of services and controlling procedures. The list below is a summary of frequently used PSTN supplementary services.

- Call forwarding unconditional/ on busy/ on no reply
- Call waiting
- 3 party conference call
- Call forwarding to individual announcement
- Do not disturb / call rejection
- Call restriction permanent/ with password
- Abbreviated dialling
- Hot line
- Multiple number with different ringing
- Calling line identification presentation
- Calling line identification restriction
- Calling line identification at call waiting
- Voice mail message waiting indication
- Anonymous call rejection
- Completion of call to busy subscriber
- Completion of call on no reply
- Selective ringing / call waiting / call forwarding / call acceptance
- Wake up call single / regular
- Carrier selection call-by-call / preselection
- Call back to last no answered call

5.2.6. ISDN Supplementary Services

5.2.6.1. ISDN Basic and Primary Access

At the digital connection from the ISDN user to the ISDN local exchange the organization of ISDN interface channels (B – speech channel, D – signalling channel) may be:

- Basic access: 2 B + D (B = 64 kbit/sec, D = 16 kbit/sec)
- Primary access: 30 B + D (B = 64 kbit/sec, D = 64 kbit/sec)

The very big set of supplementary services supported by the ISDN utilizes the DSS1 common channel signalling transmitted by the D channel and the intelligence implemented in the ISDN terminal equipment.

The ISDN basic access supports two kinds of configurations. The point-multipoint configuration enables connection of several individual equipment even with different call number like ISDN telephone, fax, and computer. The point-point configuration enables connection of ISDN PBXs. The ISDN primary access usually means point-point configuration.

5.2.6.2. User Control of ISDN Supplementary Services

Two kinds of procedures can be used for controlling the ISDN supplementary services:

- Keypad protocol: the user applies digit codes in between * and # codes like in PSTN supplementary services. The terminal equipment has no intelligence so it is not able to recognize type and status of services.
- Functional protocol: standardised messages are sent and the intelligence is shared in between the terminal equipment and the local exchange. Both parties have to know the type and status of services and have to support functions concerning the required service.

5.2.6.3. List of ISDN Supplementary Services

Standardised and individual supplementary services can be offered to ISDN subscribers. The most frequently used services standardised by ITU-T and ETSI are listed below:

- Calling Line Identification Presentation (CLIP)

- Calling Line Identification Restriction (CLIR)
- Connected Line Identification Presentation (COLP)
- Connected Line Identification Restriction (COLR)
- Terminal Portability (TP)
- User-to-user signalling (UUS)
- Closed User Group (CUG)
- Sub-addressing (SUB)
- Malicious Call Identification (MCID)
- Conference Call (CONF)
- Call Forwarding Unconditional (CFU)
- Call Forwarding on Busy (CFB)
- Call Forwarding on No Reply (CFNR)
- Call Deflection (CD)
- Call Hold (HOLD)
- Call Waiting (CW)
- Three Party Service (3PTY)
- Message Waiting Indication (MWI)
- Completion of Call to Busy Subscriber (CCBS)
- Completion of Call on No Reply (CCNR)
- Advice of Charge (AOC)

5.2.7. IN Services

In the description of IN concept is stated that the services has to be created quickly and flexibly and easily adapted to the user. However the implemented systems usually have pre-programmed services and the flexibility and user options are limited to selection of parameters of services.

The most frequently used IN services are:

- Freephone services: the call is free of charge for the caller; fees are covered by the called party.
- Shared cost services: the caller parties pay a lower level e.g. local tariff; the rest of the actual fee is covered by the called party.
- Premium rate services: the caller parties pay a higher-level fee and one part of the fees is forwarded to the called party.

The 3 services listed above have some common features. One is that a specific NDC – National Destination Code is assigned to the services, actually in Hungary the NDC for freephone is 80, for shared cost service is 40 and for premium rate service is 90. The charging principle of calls is completely different from the ordinary calls. The termination of calls might depend from the calendar, time of day, or the origin of the call.

- Virtual card: the users cover the fees of the initiated calls by his own normal telephone bill. The actual call has not charged to the line from the call is originated.
- Prepaid card: the user buys an account with a given value. The fees of the actual call are charged to this account. The card is used to identify the account. The card can be refilled.
- Virtual private network: the defined public and PBX lines can form a virtual private network. The calls within the network can be initiated by short numbers and the fees are adapted to the internal tariff schemes.
- Universal personal number: a special number is assigned to the subscriber and the subscriber can program by control function the lines where the actual calls are terminated depending of time of day, etc. The initiated calls using the account of the universal personal number are free of charge for the line from where the call is originated.

5.2.8. GSM Supplementary Services

5.2.8.1. Customized Services

An important feature of the GSM services that the terminal equipment and service provision can be assigned to a given user. The traditional PSTN/ISDN terminal however can be used by several persons who can access to the telephone set. The intelligent mobile terminal equipment can be programmed according to the individual needs of the user.

5.2.8.2. Prepaid Services

Among the GSM services very popular is the prepaid system beside the traditional subscriptions (paying the bill monthly). The fees of the calls are charged to the prepaid account. The calls can be received even in cases of 0 account. In this system the account balance have to be managed even during the call so the set of services is usually limited e.g. international roaming due to technical difficulties of real time account management.

5.2.8.3. . CAMEL - Customised Applications for Mobile network Enhanced Logic

The international standardisation has targeted this very special intelligent network function. With CAMEL the customer roaming in foreign network can sense just the same service environment which is programmed in home conditions including the non-standardised individual supplementary services like call filtering, selective call forwarding, announcements and interactive information retrieval.

5.2.8.4. GSM Supplementary Services

- Calling Line Identification Presentation (CLIP)
- Calling Line Identification Restriction (CLIR)
- Connected Line Identification Presentation (COLP)
- Connected Line Identification Restriction (COLR)
- Call Forwarding Unconditional (CFU)
- Call Forwarding on Mobile Subscriber Busy (CFB)
- Call Forwarding on No Reply (CFNRy)
- Call Forwarding on Mobile Subscriber Not Reachable (CFNRc)
- Call Deflection (CD)
- Call Hold (HOLD)
- Call Waiting (CW)
- Multi Party Service (MPTY)
- Barring of All Outgoing Calls (BAOC)
- Barring of Outgoing International Calls (BOIC)
- Barring of All Incoming Calls (BAIC)
- Barring of Incoming Calls when Roaming Outside the Home PLMN Country (BIC-Roam)
- Closed User Group (CUG)
- Advice of Charge (AOC)
- User-to-user signalling (UUS)
- kbit/sec, D = 64 kbit/sec)
- The very big set of supplementary services supported by the ISDN utilizes the DSS1 common channel signalling transmitted by the D channel and the intelligence rding on Busy (CFB)
- Call Forwarding on No Reply (CFNR)
- Call Deflection (CD)

- Call Hold (HOLD)
- Call Waiting (CW)
- Multi Party Service (MPTY)
- Barring of All Outgoing Calls (BAOC)
- Barring of Outgoing International Calls (BOIC)
- Barring of All Incoming Calls (BAIC)
- Barring of Incoming Calls when Roaming Outside the Home PLMN Country (BIC-Roam)
- Closed User Group (CUG)
- Advice of Charge (AOC)
- User-to-user signalling (UUS)

5.3. Using of the Dialog Based Service Networks for Other Purposes

Erika Madarász , author

György Takács PhD, reviewer

The traditional telephone service is a typical two-way symmetric real time dialog based service as it is described in the point 5.2. As results of telecom development, new services and applications have appeared bringing new networks or new way using of the existing networks. The real bottleneck in the introduction of new services is the network.

The global telephone network has extremely high benefit based on the well established and fully interworking system of connected networks. The system of networks developed for telephone services is suitable for quite different services then original telephony. In developed countries practically the full population has access to the telephone networks. The global telephone network is a really big asset, so its extended utilisation has good reason in economic point of view as well.

The basic ability of the telephone system to connect the termination points quickly and transfer any kind of information in between servers and users suggests introduction of new services and applications. The range of potential new applications seems to be extremely wide so we list here examples only. First of all her we sum up the network features having importance in the introducing new services and applications.

5.3.1. Dialog Oriented Networks

The telephone networks are traditionally circuit switched networks with analogue or digital access part based on wire-line or mobile technologies.

5.3.1.1. Wire-line and Mobile Networks

In the PSTN the voice band transmission possible in the frequency band 300-3400 Hz by the traditional technologies. The digital transmission and switching are spreading however users need access through cheap analogue sets. The analogue line interfaces and sets limit very often the capabilities of networks. The digital technologies like ISDN support functional developments as well. The integrated service means, that the whole system generally supports bearer and teleservices.

The PSTN/ISDN networks have both analogue and digital access parts but the switching part, transmission network part and the signalling systems are common. The interconnected global networks have different level of technology and different level of capabilities in the subnetworks, so the global network is not homogenous. This is a limit in spreading of unified services.

The mobile telephone networks are separated systems interconnected with each other and with the PSTN/ISDN world by well defined interfaces in the gateway exchanges. The first generation mobile systems were developed purely for voice communication. However the second generation digital GSM systems were developed to support several teleservices and bearer services.

5.3.1.2. Teleservices and Bearer Services

Teleservices are defined at the user interfaces but the bearer services at the network interfaces of the terminal equipment as can be seen in figure 5.3.1.:

- teleservices are telecommunication services which provide full communication in between users according to the agreed protocol of network operators including the terminal equipment functions
- bearer services are telecommunication services providing the signal transmission capabilities in between network access points

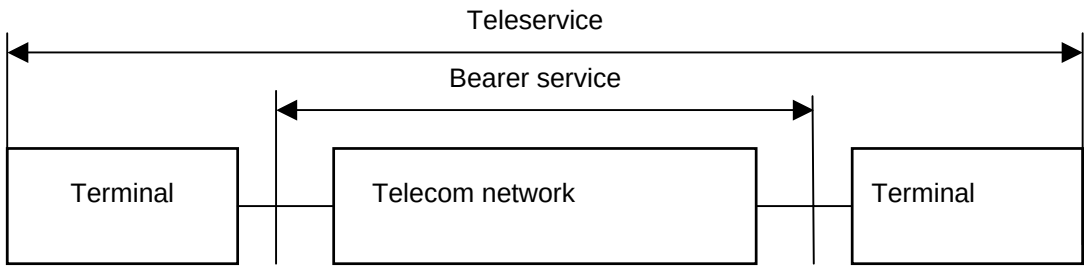


Fig 5.3.1 Definitions of teleservices and bearer services

Bearer services have not direct connection to the end users, but have importance to connect correctly terminal equipment. So concerning applications the barrier services have only secondary importance and there will be presented by some examples.

The ISDN bearer services include:

- 64 kbit/s, unrestricted, 8 kHz, structured circuit-mode bearer service,
- alternate speech and 64 kbit/s, unrestricted, 8 kHz, structured, circuit-mode bearer service,
- virtual call and permanent virtual circuit packet-mode bearer service

Teleservices are directly used by end users. The benefits of standardised teleservices are the smooth interworking of network operators, economy of scale in manufacturing and selling terminal equipment which means safe and cheap communication.

The widely spread ISDN teleservices are:

- telephony - providing two-way real time conversation on the network in between users,
- teletex – transmission of documents as teletex coded information from memory to memory on the network,
- G4 telefax - transmission of documents using Group 4 faximile coding on ISDN
- videotex – access and downloading text and graphical information extended with mailbox functions,
- 7 kHz telephone - good quality two-way real-time voice transmission using 50-7000Hz bandwidth on ISDN

The most popular standardised GSM teleservices:

- telephony
- emergency calls
- short messages – terminated in the handset (SMS MT/PP)
- short messages – initiated by handset (SMS MO/PP)
- short messages – broadcasted within the cell (SMS CB)
- alternated voice and telefax G3
- automatic telefax G3

Some details concerning the teleservices listed above are the followings.

5.3.2. Telefax

The telefax services are developed for transmission data instead of voice on the network in between the terminal equipment. The data transmission needs only one way communication and the necessary other communications are managed by the terminal equipment. The terminal equipment scan the document, the scanned signals are coded into digital form and transmitted over the network using the standardised protocols. The receiving terminal decodes the information and produces the copy of the original document (faximile).

The telefax was used in the PSTN at the beginning but had so quick success so had to be implemented in ISDN and GSM as well. The key issue of the success was that developments in the network were not needed to introduce telefax services but only connection of new terminals to the existing network. The faxed documents can be handled as real copies of the original documents. The business and administrative world have accepted as real documents. Faxes are faster and cheaper than postal delivery of documents.

Interworking of telefax terminals are based on standardised protocols. The G3 (Group3) terminals use analogue PSTN line interfaces. The digital codes are presented on the transmission lines in the form of modulated audio frequency signals. The G4 terminals transmit the digital signal more efficient mode on ISDN to communicate with an other G4 terminal. The G4 terminals can forward faxes in G3 mode as well if the destination terminal can use only G3 protocols.

The telefax teleservice standardised in GSM support G3 protocols only. In the GSM systems the automatic telefax is a separated teleservice because the voice transmission has compressed speech coding which does not support fax code transmission. The air interface has no speech coding in automatic telefax mode. The alternative speech/fax mode is a special one and the coding can be controlled during a call.

The telefax transmission has the following steps:

- Call setup – like as normal calls using the telephone number of the called terminal.

- Negotiation of modem transmission mode – using standardised protocols the terminals fix the mode and the appropriate transmission speed adapted to the actual network performance.
- Data transmission – the scanned information is transmitted using the agreed mode. In cases of changes in the network performance during the call the data rate can be adapted.
- Acknowledgement – feedback data from the receiving terminal on the successful transmission or on the errors. The results are indicated or printed on the sending terminals.

5.3.3. Voice Mail

The basic function of voice mail services is recording messages. This is a tipical on-way non real time service to provide recorded messages from the caller users to the called users.

The voice mail system is a centralised message recorder. The callers leave messages and the called users can listen them, and message delete function and archivation are also available. The mailboxes can be subscribed and the complex system can manage thousand of boxes. The boxes can be assigned to existing lines or might be individual boxes.

Messages can be recorded in the mailbox assigned to a given telephone number if the conditions defined by the user are valid, e.g. busy line, no answer. Such boxes have no own phone numbers, calls and messages are identified on the basis of called number. An important task of the voice mail system to inform the user on the received messages. The first step is transferring indication on recorded messages from the mailbox system to the telephone switch. The second step is to transfer the message waiting information from the telephone exchange to the subscriber:

- Initiating calls to inform the users. This solution is simple but generates traffic load with low chance of success.
- Modified dial tone. This solution needs modifications in the telephone exchanges and the user can detect waiting messages only cases of picking up the handset.
- Display waiting messages with a lamp on the telephone sets. This solution needs considerable extension in the signalling system but can be combined with other services like display of calling number or SMS.

Listening messages can be managed easily from the assigned telephone sets but remote retrieval can be possible as well.

The individual mailbox has its own phone number and the mailbox system is connected immediately to announce that ready to record messages. It is impossible to indicate waiting messages. The recorded messages can be listened from any telephone set.

The fax mailbox service can store faxes. Its working principles are similar to voice mail systems and voice mail and fax mail can be integrated. The fax mailbox service might be very useful to manage intensive and bursty fax traffic. Circular faxes can be managed easily by fax mailbox systems.

5.3.3.1. Application of voice mail in different telephone systems

The voice mail systems were introduced in PBX systems in the first period. A closed and limited network can be managed relatively easily. The next successful field of its application was in the mobile telephone systems. The switched-off handsets and the limited coverage suggest the application of mail boxes. The implementation has been relative easy because of the direct support of the signalling system. The indication of received messages has several excellent solutions e.g. by SMS.

In the ISDN/PSTN networks the message recording appeared in the first period as a special terminal equipment combining telephone sets with a small magnetic tape recorder. The real voice mail systems were introduced in mid nineties. The indication of waiting messages are listed above.

A new line in the service development is the universal messaging where messages received in different form and handled by the same way. The messages can be converted into other form e.g. automatic reading of e-mails, forwarding voice and fax messages by e-mail, forwarding the header of e-mail by SMS.

5.3.4. SMS – Short Message Service

The SMS seems to be the most successful non-voice service in the GSM networks. The SMS enables sending text messages from one handset to an other handset. The sending a message has two phase:

- Message transfer from handset to the SMS service system (MO/PP – Mobile Originated/Point to Point). The message might be maximum 160 text character. The destination might be a mobile subscriber defined by its call number. Special agreement of network operators can support forwarding the message to a pager or to a PSTN/ISDN terminal.
- Message transfer from the SMS service system to the handset. (MT/PP – Mobile Terminated/Point to Point).The origin of messages are generally handsets but operators and other systems (e. g. Internet, unified message systems) can generate messages as well. Acknowledgement on delivery of SMS can be managed.

The third GSM teleservice concerning the SMS is the message broadcasting within cells (CB – Cell Broadcast). This teleservice enables sending messages containing maximum 93 characters. Concatenating such messages longer message can be composed. All of subscribers in the area of the cell can receive the broadcasted message. The handsets can filter the messages having already been received on the basis of message identifier. The message broadcasting has importance in propagation urgent local information.

The SMS applications have continuous development. Some examples are listed below:

- Personal messages between mobile subscribers – this is the most popular application of SMS. Their number is comparable to the voice calls.
- Subscription for regular information provision using a content provider e.g. exchange rates, stock rates, lottery results
- Case-by-case SMS messaging based on pre-defined events e.g. transaction with credit card.

The success of SMS in the GSM networks raised the demand of PSTN users to apply similar system in the network. The necessary technical developments have started. The critical part is the signalling system supporting fully the SMS. The ISDN seems to be easier task due to its signalling features. The analogue system needs further development of calling number display protocols and functions.

5.3.5. Internet Access

The PSTN/ISDN system offer an access to the services of the global Internet systems as well. The Internet service providers (ISP) have to install gateways for interconnection of ISDN/PSTN and Internet systems. The gateways handle IP addresses on the Internet side interfaces and telephone numbering schemes on the telephone side interfaces. The connection to the Internet can be set up by dialling phone number to reach the gateway and the data transfer uses modems managing line coding.

The ISP has service agreement with the customers defining the conditions to use Internet services. The ISP manage the authorisation and authentication procedures. The billing system store the usage records and assemble the bills. Internet access through PSTN has no limitation in using services e. g. www, e-mail, FTP, CHAT but the data rate on the PSTN lines can limit the download speed.

5.3.6. WAP

The WAP (Wireless Application Protocol) is a special browser with adapted functions to mobile networks and content is adapted as well. The air interface limits the downloading speed and the displays of handsets are usually small so the WAP based information management need specific structure and presentation.

The content on WAP based portals are growing and the GSM service providers usually develop their own portals.

5.3.7. Telephone based information retrieval and administration

Information and entertainment

Broadcasting of sound programs using wireline systems seems to be as old as public telephony. Even nowadays a big sort of information can be received by telephone systems.

Services can be classified on the character of information.

- Public interest information (e. g. speaking clock, weather forecast, time tables, travelling information, mental aids)
- Commercial information (e. g. exchange rates, tax tips)
- Entertainment (e. g. games, tales horoscope)

The mode of information provision might be pre-recorded speech, selection from menu points with defined structures, leaving messages, real personal contact with operator or a combined form.

Another classification might be on the basis of tariff schemes.

- Some information can be retrieved charging by normal telephone tariff.
- Other kind of information available on case based charging (e.g. directory services)
- Special schemes can be organised also like free calls, shared rate charging, premium rate charging.

Call centres, Contact centres

Customer services can be supported high level with new applications of telephone systems. The full availability of telephones, the integration with information technologies give nice possibility to serve customers quickly, and with high comfort and reliability. The competition in services also accelerate the application of call centres in several fields of the life.

The main function of call centres to receive phone calls and forward the calls to the relevant person to serve the customer providing information on products or services, receive the data of customert, receive orders…. Commercial and service companies where the customer contact are crucial (e. g. banks, assurance companies, telecom companies, airline companies, department stores, ticket offices) can utilise the benefit of call centres. Call centres manage the front line activities of such companies so correct handling of the incoming calls have high importance in the customer satisfaction. This goal can be reached by optimal combination of high technology and the well trained staff. Call centres are good tools in managing and monitoring of the business activities.

The technology applied in the call centres has continuous development in integration of telecom and IT systems to have better and better adaptation to administrative processes and customer handling. The Automatic call Distributor (ACD) and Interactive Voice Response (IVR) systems enable the correct and quick

handling of customer calls and the further development can manage beside phone calls telefax and Internet contacts. The Internet call centre can send e-mail from the customer with a singly click on the home page or initiate a phone call. The VoIP features in the Internet computer can initiate VoIP calls by a singly click. The customer and the agent can see synchronised web pages and fill the ordering form with parallel verbal discussion. The continuous help of agent can result successful ordering or electronic transactions.

Contact Centre is a new generation of call centres supporting the professional and well-qualified agent staff in customer oriented activities to answer the requests of customer in one or a few steps. The key elements of contact centres are the Customer Relationship Management (CRM) systems to integrate the call handling with business processes and databases to support the agents.

5.3.8. Other Applications

The possibilities in dialog based service networks for innovative applications have no strict limits. The summary below lists only a few applications in a short form.

Conference Call

Conference calls connect more than two terminal equipment for multilateral communications and a special technical solution is needed to implement conference services.

Setting up conference calls can be managed by one subscriber or by an authorized operator to enter the conference members into the conference call step-by-step.

Other solution for conference calls when conference member can enter into the service dialling a pre-defined phone number. This type of conference call can be used in two forms. Closed user application is possible if the phone number of the conference call is not published only the selected persons know it and call the conference system in a pre-agreed time. Public service can be provided with publishing the conference call number and the topic of conference. Such chatting applications are more and more popular and the most favourite topic is the partner hunting.

Beyond voice conferences video conferences can be provided on ISDN. Video conference terminals and controlling functions have to be installed on the spots of the conferences. Companies and organizations utilize the benefit of conference calls for discussions reducing travel cost and time of staff.

Location oriented services

One of the special content provision is when the exact location of the user is an important selection parameter in information. The position of customer is determined automatically based on location server function. In PSTN/ISDN the location function can be determined by IN functionality to detect the geographic location from the origin of call. In the GSM system based on base stations the geographic locations are available.

Big department stores and chains of service points can transfer the customer orders to the nearest service point (e.g. pizza service) both in wired and mobile networks and location specific information can be also provided like local transport, local weather, stores, restaurants.

Televoting

Televoting services can collect and count of votes produced by phone calls. The actual number of calls can be monitored and final result also. Selection among possible choices might be calling different numbers.

Televoting can be used to select the predefined order number of callers and forward the selected caller to operator. Separated counting of different geographic origin can be managed also. Rejection of repeated votes is also possible.

Donation by calls

The telephone network can be used for charity purposes also. The published phone numbers assigned to charity usually are charged with premium rate. The given portion of fees are transferred to the organization ordering the service.

Burglar alarm services

The telephone system is very suitable to transmit the alarm signals in case of burglary, fire etc. to disaster management centres or the pre-defined subscriber. The alarm services are popular among companies, offices and residential users.

5.4. Content services

Házkötő Béla, author

Sándor Stefler, reviewer

During cultural history of the mankind information had always a great significance. At first conveyance of information was only verbal expanded with several signals or marks. There was a hard disadvantage, that this information could by developed only in the moment of conveyance, but in the next moment it disappeared and could be not reproduced only with loss or distortion. Therefore mode of storage was soon be looked for and found. At the very start primitive figures and signals scratched or engraved on rocks, on woods, or on bone spread information for some limited persons or little groups. Later on discovery of different sorts of writing resulted further and durable possibilities. Development of various writing technologies (by characters, pictures, runes etc.) on isolated districts of the World at about equal time proved the stark claim for the storage of information. The invention of typography has brought a significant change. It has given the possibility for reproducing information in theoretical by unlimited amount, quite cheaply and suitable for distribution in big masses. Newspapers have appeared. By this time preparation of the content (editorial office) and distribution (publisher) where separated, and representational carriers (e.g. encyclopaedias) appeared. In practice theoretical grounds of information's content services have been developed.

Possibility of the explosion-like evolution was given by the technical development, electrical engineering, and later by the development of electronics and propagation of telecommunication. At first the telegraph was only a technical expedient for journalism, but the telephone provided shortly original possibilities.

Just some years after inventing the telephone in 1881 Tivadar Puskas presented the „Telefonhírmondó" (telephonograf) on the world exhibition, in Paris, which became the first program distribution system on the world after its premier in Budapest in 1882. It was broadcasting varied programs day to day 14,5 hours long for several thousands of subscribers. After about 30 years of functioning it gave up its place for a new „pretender": the radio. From that time the „milestones" of the

evolution followed each other faster and faster. After about 10 years started the first wing-beats of the TV which, first with black and white, then with colour technology - created a new form in the media. In the meantime the computer has appeared and informatics started a stormy evolution, too. Diverse forms of telecommunication with greater and greater output have been evolved and the two fields began converge. A world-wide computer network has been built up and the Internet captured the world by a never learnt speed providing invaluable perspectives for the mankind. A new period of the informational revolution has arrived: accessibility to information has become a fundamental right of the human being, and establishing of the information's society has started.

5.4.1. Content

In this chapter services providing information content are surveyed. The content as the subject of certain services represents some similarity but it differs, too, dependent upon technologies and other factors. The free flow of information and access to information are parts of fundamental human rights. [5.4.1]

Information can contain materials that are illegal or harm interests, moral, or religious conviction of other people or groups of people and may be harmful for development of children and minors. The European Convention on Human Rights authorises governments to take bans and restrictions concerning undesirable contents without violation of the human right.

Undesirable contents can be divided into two major parts: illegal and legal, but harmful especially for minors. Estimation of the first group is unambiguous and homogeneous; the second's may vary depending on the country.

Representation of violence in media.

Let us look over some possible aspects of the relationship of media and violence concerning the complexity of the problem.

The public and its access to the electronic media

a.)Television

- free access (unencrypted)
- fee-paying access (encrypted)

- "professional" access (medical pay TV)
- interactive television (using for example video games, CD Rom or Internet)
- programming time (children's programming time / prime time / programming time after watershed).

b.) Other

- Internet
- video (free or conditional access).

Types of programmes

a.) Television programmes: news, current affairs, documentaries, science programmes reality shows, light entertainment, music, video- clips, game-shows, contests, sport, religion, children's programmes, films, drama, advertising, teleshopping, trailers.

b.)Radio programmes: news, current affairs, light entertainment, music, sport, religion, youth, and advertising.

c.) Other: video-cassettes, trailers video games, multimedia.

Acts of violence portrayed:

- physical violence, sexual violence, psychological violence, verbal violence, implied violence, threats, act in itself (e.g.: physical aggression), result only (e.g.: injury or death, material damage), act and result.

Context of portrayal of violence:

- information, education, awareness-raising (charity), artistic expression, entertainment, social criticism, irony, humour, audience attraction/sensational, unintentional.

Form in which violence is portrayed:

- realistic, naturalistic, hedonistic, esthetic, aggressive, raw material, picture and comment/value judgments, positive/negative (violent act of the hero/anti-hero). [5.4.2]

Protection of minors and human dignity

Human dignity and protection of minors against harmful influences are fundamentally important parts in providing audio-visual and information services. This kind of materials can be divided in two parts:

1. contents, publication of which is illegal and as a crime, it is punishable, e.g. child pornography, violent pornography, extreme gratuitous violence and incitement to racial or other hatred, discrimination. The most countries have prohibitions on this kinds of materials on producing, distributing, importing and

advertising, too. Like it, are prohibited publication such as material that is obscene, contrary to sound morals or indecent.

2. contents, lawfully publicized to adults are available also to children, but may be harmful to the latter's mental and moral development. There are some problems on this field. Globalisation of services, audio-visual services without frontiers, very speedy growing of the Internet reveal the differences of diverse cultural and moral traditions in certain countries which influence estimation of concrete forms and identification of undesirable contents. The newer and newer technical facilities also change the character of the services, too. Previously the choice of a subscriber was limited to the question „watch or not to watch TV?." Today the choice has grown very wide, even with utilisation of video-on demand (VOD) or similar possibilities and every user may become a potential program-editor. [5.4.3]

According to the former mentioned facts, different technical and administrative means may play an important role in protection of children. One of theme may be the controlled access. Simple methods in the broadcasting are sending programs late in the evenings or labelling the programs with conventional symbols according to content's categories. Both methods shift the responsibility upon parent's control.

New services generally include a restricted access for children e.g. by the existence of subscription, by using an identifications code or simply by the contact between the provider and the subscriber.

Possibilities of protection

Canada has been a pioneer in protection of minors. In the 1990s launched a strategy for fundamental frameworks of protection:

- collaboration based on the principle that violence on television is a cause of psychological problems among children;
- protection of children and not censorship of adults;
- concentration on gratuitous or idealized violence without involving eroticism or other moral considerations;
- involvement of all those concerned: broadcasters, advertisers, producers, parents, teachers and specialists in mental health;
- adoption of a dual perspective, one short-term and one long-term;

The most important things to be done are the follows:

- codes of conduct worked out with the industry;
- classification of programmes;
- anti-violence or „V" chip;

- information campaign to increase public awareness, and media education programme. [5.4.3]

Note: the „V" chip is a technical device incorporated in a television set, with it is impossible to view programmes, classification by a provider's code, if it is not derestricted (for example by parents). It is possible to restrict different kind of programmes with different codes. Application of this method grows slowly.

Canadian model constitutes the basis of protection today all over the World.

The simplest method is the controlled access. On field of broadcasting the simplest methods are sending programs late in the evenings or labelling the programs by conventional symbols according to their content's categories. Both methods throw the responsibility upon parent's control.

Control

In Hungary propagation of programs concerning radio and TV Act No. 1 of 1996 on radio and television services controls broadcasting fundamentally.

„*Article 3 (1) In the Republic of Hungary broadcasting may, within the framework of this Act, be freely exercised, information and opinion may freely transmitted through broadcasting, and Hungarian as well as foreign programmes intended for public reception, may be freely received. The broadcaster, within the framework of this Act*, *defines the content of broadcasting by itself and is responsible for it.*

(2) The broadcaster shall respect the constitutional order of the Republic of Hungary, its activity may not violate human rights or encourage hatred against individuals, genders, peoples, nations, national, ethnic, linguistic and other minorities, and church or religious groups.

(3) Broadcasting may not aim, openly or implicitly, at insulting or segregating any minority or majority. Nor may it *present or judge them from racial points of view*."

„Article 5 (3) No images or sounds presenting violent conduct as exemplary may be broadcast in programmes for minors.

(*4) Programmes that may have an adverse effect upon the development of the personality of minors, particularly programmes presenting the gratuitous use of violence as an exemplary conduct and programmes presenting gratuitous sex may*

only be broadcast between 11:00 p.m. and 5:00 a.m. The attention of the audience shall be drawn to the nature of such programmes prior to broadcasting.

(5) It is forbidden to broadcast any programmes that could have a seriously adverse effect upon the development of the personality of minors."

By this Hungary satisfied the European agreement concerning „television without frontiers", Strasbourg 5. May 1989 [5.4.5], which was pronounced by Act No. 49 of 1998 [5.4.6].

5.4.2. Services

Nowadays we are witnesses of the enormous and speedy transformation of the content services. The development of digital technologies, the continuous appearance of newer and newer services and the convergence of telecommunication and informatics fundamentally change accustomed and existing structures. Therefore statements, according to actual regulations, are not durable. The legal environment is forced to continual evolution, too. Programme transmission as an independent service will presumably be left out from programme delivery and this function will be changed by simple leased line services. Cable networks are prior a big development especially after liberalisation. Interactive wideband networks of a good quality being under construction will provide an enormous assortment of content services for the subscribers. With reliably solution of downlink channels satellite programme distribution will also find place in the competition of the technologies. Developed services „tailored" for mobile communication supported by UMTS Universal Mobile Telecommunications System (UMTS) will bring hardy estimable changes in the application of the information.

What can this „information supermarket consist of"? By a vision, for example of the followings:

Information, i.e.

- news, summaries, business and finance, politics, culture and entertainment, sport, gambling, etc.
- financial services: stocks, stock-market, current rates, charge accounts, credit cards, money transfers, statement of accounts, etc
- localised services: restaurants, theatre-cinema-concerts, inquiry offices, correct time, pharmacies, supplication, weather forecast, etc.

- commerce: categorised (property, car, employment), shopping (daily, actions), tickets, etc.
- travel: travelling information, orientation, time-tables, hotels, holiday-offers,
- private interest: Internet, computer technique, car, antiquities, etc.

Communication, i.e.

- messages: SMS, postcards, multimedia-messages,
- E-mail: sending/receiving, dictaphone,
- group communication

Power, i.e.

- organisation: task-lists, address lists, agenda, memos,
- personal assistance: registers, translation, etc.
- resources: calculator, dictionary, translator, etc.
- mixed: control of domestic machines, automatics, identification of persons and places, home networks, etc.

Entertainment, i.e.

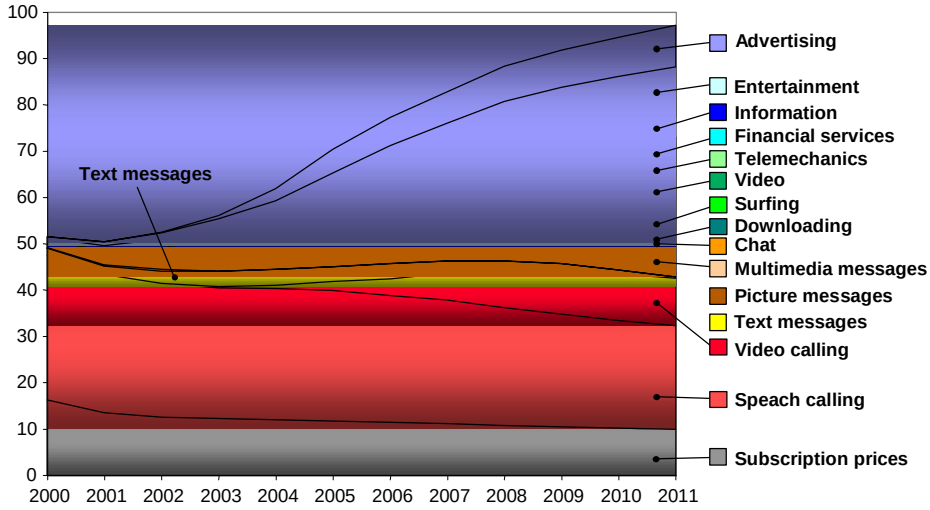
- music, TV, lifestyle (hobby, mode, company), stories, chat, pictures, games, astrology, biorhythm, first experieces of somebody, etc.

Source: Nokia

There should be an estimation of a service provider concerning the trend of an average monthly receipts pro subscriber in the next 10 years.

Average revenue per user (ARPU)

Estimation



Source: Nokia

Broadcasting

„(32) Free-to-air broadcasting: one-way radio communication procedure carried out whit a terrestrial or satellite system and intended for a theoretically unlimited number of users with suitable receiving equipment, for the transmission of sounds, images or signals of other nature." [5.4.4]

Detailed rules of radio and TV broadcasting can be found in the Act No. 1 of 1996. [5.4.4]

Special forms of broadcasting are the public service broadcasting and the public broadcasting. The aim is to supply the general public with public service programmes with respect of the dignity and other interests of the national, ethnical, lingual and other minorities without injuring other nation's dignity. It provides regular, overall, unprejudiced, authentic and correct information about local and foreign events, supplies information, counting on public interest with typical and contrary opinions. It attends to make communications of general interest public. Its mission is to satisfy interest of the public variegated and on a high level. It gives curious attention to present and preserve cultural values included also religious, national, and ethnical and other minority cultures. Additional mission is to send programmes, in prime time that serve development and knowledge of minors and spread important information for disadvantageous groups especially regarding to services that introduce to rights of minors serve their defence and give information on available services.

From the above-mentioned follows the need of the economical, political and other independence of public broadcasters. This is suggested both by national and international right. The most important question is: how to ensure the economical independence.

The funding schemes for public service broadcasting are fundamentally different from the funding schemes of other services of general interest (such as „universal service funds" in telecommunications).

There is a consensus in Europe that public service broadcasting needs an appropriate, secure funding framework, and that public funding is an integral part of public service broadcasting systems.

What are the terms „public” and „commercial” financing? Revenue, which comes from the state budget or other public funds, or from mandatory fees paid by viewers and listeners (broadcasting license fees), is generally referred to as "public revenue". In contrast, income from contractual transactions on the market (advertising, sponsorship, programme sales, etc.) is generally referred to as "commercial revenue".

In Europe and in Hungary too, the mixed model is used, where a combination of public and commercial revenue is used to found the public service activities. But we must to allow that "commercial revenue" may contribute to the funding of public services, but "public revenue", which has been granted for the fulfillment of public services, must not be used to fund services outside the public service remit, they must operate a strict separation of accounts, as it is prescribed in the recent European Directive on financial transparency.

The predominance of mixed funding in Europe may be explained by the following consideration:

- In many countries, a single source founding would not provide sufficient revenue to guarantee the fulfillment of the public service remit;
- Reliance on one particular source of funding creates dependencies, which run the risk of undermining the independence of the broadcaster and the effective fulfillment of the remit;
- Mixed funding systems may be considered more robust in a rapidly changing environment, where certain sources may suddenly dry up while others grow and new ones emerge.

The funding of public service broadcasting is more than a financial issue. Although the rule „Who pays commands” cannot automatically be applied to broadcasting, the way funding is provided is likely to influence broadcasting output. Broadcasting journalists and managers are well aware of who ultimately pays their salaries: viewers and listeners, the government, or the advertising industry. Apart from the psychological links and affinities there may also be hard economic pressures linked to particular sources of funding. For example, the more dependent a broadcasting organization is on advertising/sponsorship revenue the more it will be under pressure to achieve high audience rating with regard to those sections of the population which are of primary interest to advertisers. This leads to a typical „deformation” of programme schedules, with preference being giving to popular or

relatively cheap mainstream programmes, avoiding difficult, controversial or experimental programmes, and neglecting the interests of certain age-groups and minorities. The result is „commercialization”.

On the other hand, where a broadcaster has to rely primarily on direct State funding, there is the risk that the public authorities or political parties will use this „leverage” to gain influence over editorial policy. Even without such interference, broadcasting organizations may be inclined to „please” bodies, which have decision-making power over their funding. They may be tempted to hold back „negative” information, avoid programmes which are critical of the government, etc.

The funding schemes by different sources have one by one any advantages and disadvantages. Ratio of different funding schemes, which is chosen suitably, is optimized without risk of independence and is adjustable effortlessly to different nature of countries.

In Hungary foundations, managed by advisory boards, ensure function of public broadcasting.

The National Radio and Television Body govern the broadcasting-related rules. The most important responsibilities of the Body are the followings:

- the Body fulfils the responsibility of inviting tenders for broadcasting rights and satellite channels provided for the purpose of Government-controlled broadcasting, and it is also responsible for accessing the bids that are received,
- the Body fulfils the supervisory and inspection responsibilities stipulated by the Act,
- the Body operates a Complaints Committee for investigating case reports,
- the Body operates a programme monitoring and analysis service.

All questions related with legal status, function of National Radio and Television Body, foundations and his advisory boards together with subvention system of broadcasting can be found in Act I of 1996. [5.4.4]

Nation-wide and regional broadcasting of radio and television programmes may be provided only by a concession company or by an economic organisation established for this purpose. [5.4.8]

Program transmission

*„(29) Programme transmission: the simultaneous transmission, without any alteration of the contents, of signals produced by a broadcaster by means of a wire (cable) network or a terrestrial or satellite non-broadcasting radio communication system to radio and tele*vision broadcasting stations or programme distribution networks.” [5.4.4]

Service classification of programme transmission in Hungary corresponds with broadcasting. [5.4.8] After liberalisation of the communication it will probably leave off as a separate service and will be reckoned among the leased line services.

Programme distribution

„(26) Programme distribution: the simultaneous transmission of signals produced by a broadcaster without any alteration by means of a wire (cable) network; or by a non-free-air radio telecommunication system from the broadcaster’s premises, or from the end point of the program transmission network, by inserting a distinct organization, to the receiving equipment of entitled users, with the exception of transmitting signals by a network suited to connect less than ten receivers. A cable system within the boundaries of a real estate is not considered to be programme distribution.” *[5.4.4]*

Programme distribution is the most hopeful service of all services, providing access to information already in these days but especially in the future. In consequence of the technical development services utilizing possibilities of the digital technique has become realty, soonest evidently on TV-program distribution networks. In favour of these services two basic requirements must be fulfilled: creating of downlink channels and selective accessing of subscribers. The first requirement is needed for the provision of interactive services (they have been just started). The second has already been realised on many places for forwarding packs of programs according to claims of subscribers having different interests or solvency (or according to simply business interests of service providers).

The most important programme distribution services are the followings:

- distribution of analogue TV and radio programmes. This is the classical service of programme distribution.

- digital broadcasting. It is creation of programmes in a new form, in which the content (sound, picture, and other information) appears as digital data during the production, transmission and processing, as well. Compression of data stream by a special algorithm greatly narrows the bandwidth needed for its transmission. Digital technique in itself does not mean a great change from the side of the programme distribution service, significant difference is the claim to interactivity, which is linked with the most important attribution of the digital TV: the electronic programme guide.
- pay-per-view (PPV) service offers viewers the possibility of selection from the contents of a broadcasting schedule who have to pay only for what they actually view. These are broadcasting services whose distribution is triggered by a universally available service. Wanted programmes may be accessible by viewers possessing the requisite receiving equipment. The viewer's decision is technically confined according to the portions of the encrypted programmes he is able to decode and to view actually.
- near video-on-demand (NVOD): the principle of NVOD is similar to PPV, but the selection of the consumer is extended by distributing the material in parallel at short intervals.
- video-on-demand (VOD) offers consumers a whole range of transactional services from films on demand via tele-banking to tele-shopping. Consumers can make their choice from a catalogue of permanently available programmes. As a truly interactive service, the chosen telecommunications service will be triggered to the viewer's individual connection.
- videotext: the most typical service is the Teletel network having been launched in France in 1984 and being today the most highly developed videotext system in the world. Over two-thirds of the French population has access to the network via a specific terminal (Minitel). It has been started as an "on line telephone-directory", but now it offers some 25 000 different services, many of them transactional.
- Internet: the bigger bandwidth of programme distribution networks offers possibility to high-speed Internet access. Later on there will be the possibility to voice-telephone via Internet, too.

Video on demand

As described, interactive services give possibility for subscriber to choose a program or other content attending as video-signal, that he would like to use. By a definition: [5.4.9]

Video On Demand (VOD): A VOD application provides residential users with the ability to select among a catalogue of pre-recorded programmes (films, news bulletins, sport events, music clips, documentaries, previews, etc.) to receive the chosen programme on a TV set and navigate through it, using control commands.

When considering services such as VOD, a number of roles can be identified:

customer: Customers may rent or own the device that allows access to the different services: typically the device will consist of a television set and a set-top-box, or a Personal Computer (PC), or a work-station. Customers may get access to a service either by subscription, or on per call basis, or even by means of a combination of the two.

service provider: Service providers own and control a number of video servers from which the programmes are distributed to the customers. Service providers may be contacted by the customers directly, or through a broker. Service providers are related to content providers from which they buy the rights to distribute a certain programme. Service providers also have the responsibility of updating the video servers. Service providers may have relationship with one or more brokers.

brokers: Brokers offer a customer an access service to different service providers with which they have an agreement. They allow the Customer to choose among different service providers based on the specific user request on a call by call basis. Brokers have no direct relationship with the content providers.

content provider: Content providers own certain programmes and are able to sell the rights of distribution to one or more service providers.

network provider: The network provider offers the communication support (infrastructure) to all the parties mentioned above. It is assumed that communication will always take place through the network. It is important to note that the role of the network provider is infrastructural in the information service provision, while the other roles are of structural type. This implies that the network provider is not part of the direct chain that links the information producer (referred to as "content provider") to the information consumer (referred to as "customer"). It is expected that the network provider will provide fair and equal access to all services providers. [5.4.9]

World Wide Web (www)

The World Wide Web is a system of Internet servers that supports hypertext to access several Internet protocols on a single interface. The World Wide Web is often abbreviated as the Web, WWW, or W3.

The World Wide Web was developed in 1989 by a scientist of CERN (European Particle Physics Lab) Tim Berners-Lee. The initial purpose of the Web

was to use networked hypertext to facilitate communication among its member, than it was developed rapidly, his application grew and began to incorporate graphics, video and sound. In recent years, the use of the Web has now reached global proportions helping with this considerably to spread of Internet.

Many protocols are accessible on the Internet. The major protocols are as follows:

- E-mail (SMTP; Simple Mail Transport Protocol)
- Telnet (Telnet Protocol)
- FTP (File Transfer Protocol)
- Usenet (NNTP; Network News Transfer Protocol)
- VoIP (Voice over Internet Protocol)
- HTTP (HyperText Transfer Protocol),.

The Web provides a single interface for accessing all these protocols. This creates a convenient and user-friendly environment. Because of this, the Web's ability to work with multimedia and advanced programming languages, the Web is the fastest-growing component of Internet.

A HyperText document containing words that connect to other documents. The words are the links, which the user can select. A single hypertext document can contain links to many documents. In the context of the Web, words or graphics may serve as links to other documents, images, video and sounds. Links may not follow a logical path. The hypertext is a standardized protocol. The application of hypertext is supported by much software.

The World Wide Web contains files, called pages, containing information and additional links. Access to pages may be accomplished by:

- entering an address directly
- browsing through pages and selecting links to move from one page to another
- searching through subject directories linked to organized collections of pages
- entering a search statement at a search-engine to retrieve pages on the topic of your choice.

To find a page is possible using its address. At the beginning these addresses named URL's (Universal Resource Locator) were numerical codes translated by DNS (Internet Domain Name System). Since numeric strings are difficult for humans to use, they were changed by alphanumeric addresses. This is the format of the URL:

- protocol://host.domain/filename

The domain-name can be a multi-level one, in the beginning the second level contained the type of institution, for example .com→commercial, .edu→educational, .gov→government, etcetera. They are used now mostly in USA. In other countries of the World the two-letter country codes according to ISO 3166 are used.

To access the Web you must use a Web browser. This is a software, that allows users to access and navigate the Web. There are two types of browsers: graphical browser (the best known are Netscape and Microsoft Internet Explorer) and text browser (Lynx). Lynx is now available only for special application.

Browsers are being developed in interest of utilisation of newer and newer applications and possibilities. Little programs (applets) provide possibility to insert animated pictures in pages, multimedia applications represent of audio and video materials, real-time programs already realise Internet-TV and -radio.

Beside the problems discussed in the first part of this chapter, contents founded in the Internet propose possibility of abuse. If we set a glance over the different sorts of problems, we can see the list very extended. It is as follows:

- national security (instructions on bomb-making, -illegal drug production, terrorist activities;
- protection of minors (abusive forms of marketing, violence, pornography);
- protection of human dignity (incitement to racial hatred or racial discrimination);
- economic security (fraud, instruction of pirating credit cards);
- information security (malicious hacking);
- protection of privacy (unauthorized communication of personal data, electronic harassment);
- protection of reputation (libel, unlawful comparative advertising);
- intellectual property (unauthorized distribution of copyrighted works, e.g. software or music). [5.4.10]

Defence in this case is more difficult than in the case of other services, because the World Wide Web is global. Difficulties are generating from the „freedom" of the Internet, than to banish a given material is practically impossible, because a kind of „solidarity" prevents it. There have been and are today some attempts to control the Internet, but a truly successful model is not known. There has been spoken about self-control up to the present time. Because of convergence however,

this branch penetrates into areas that are controlled particularly. More and more outlines of complex control have been developed. Areas are to be determined where central control is possible and desirable and where self-control is needed. The first concrete results probably will appear in controlling of the E-signing and the E-commerce.

Portals

Growth of volume of information made increasingly difficult to find a theme or material from any source. A claim has arisen in a short time to facilitate the search, to sort the content by theme or by other points of view. For this purpose portals were created. Audiotext services using also menu systems for orientation, function as portals, but the most characteristic portals are located in Internet in consequence of much bigger information content.

Portals have been developed, too. First portals were intended to gleaning and accumulating information. They were organised firstly around the big browsers only for simply search (horizontal portal), then aggregated the different personal materials (vertical portal). Currently different services can also be found in the offer; in the future even software will have run by portals. There is a custom to call these development stations (like other devices) “generations” (Gen 0, Gen 1, 2 etc.).

Typical Internet portal

Different technologies require different portals because of the variety of information. Therefore also with the appearance of mobile (wap) and palm-top portals may be reckoned. Today we can speak about global portals serving different technologies according to the above-mentioned facts.

A special field of portals is the voice portal. It can make access to Internet in the form of a talk. Its technical realisation has been given by the large-scale development of speech recognition and by the speech-processing techniques. It was born also in the USA as near by all Internet applications. The idea originates from the recognition, that about 250 million computers are connected with Internet (in 2000), but about 1,3 billion phone-lines are used all over the world! Some experts estimate the growing rate of the voice-portal market 54% in next 6 years and 7 billion USD revenue with 45 million subscribers in USA (2005).

Audiotex

Audiotext service has been got known as a „value added telecommunication service". Added value of the service, accessed with a call number on normal phone-line (reserved in a number block set apart for this purpose) is given by the accessible information. Information may be either conserved or live speech and combination. The service may be interactive or not. In case of an interactive service the user signals the code of the needed direction by pushing buttons of his telephone, according to the commands or information he has got. Information may be accessible by a menu-system. This system may be simple or multi-level. Use of the service does not need a special terminal.

This service provides the most different information e.g.: weather forecast, timetable, evening tale, speaking clock, cure of souls etc. Services with erotic content are widespread. Because of this all problems analysed above have been arisen. Therefore service providers give possibility for certain or selective prohibition in calling of audiotext services in accordance of the subscriber. There is growing self-regulation of audiotext suppliers in many countries both on national and on international level.

In general the telecommunications service provider collects the charge of service and shares it with the audiotext service provider according to their contract, but it is possible to use a pre-paid card or a subscription.

References

[5.4.1]: Az Emberi Jogok Európai Konvenciója (ETS No 05: Convention for the Protection of Human Rights and Fundamental Freedoms

[5.4.2]: Council of Europe Rec. No R(97)19 on the Portrayal of Violence in the Electronic Media

[5.4.3]: COM(96)483 Green Paper on the Protection of Minors and Human Dignity in Audiovisual and Information Services

[5.4.4]: Act I of 1996 on Radio and Television Broadcasting

[5.4.5]: ETS No 132: European Convention on Transfrontier Television

[5.4.6]: Act XLIX of 1998 on Promulgation of Text of the European Convention on Transfrontier Television, Strasbourg, 05.05.1989.

[5.4.7]: EBU-UER: The Funding of Public Service Broadcasting 9.11.2000

[5.4.8]: Act LXXII of 1992 on Telecommunications

[5.4.9]: ETSI ETR 262: Broadband Integrated Services Digital Network (B-ISDN); Asynchronous Transfer Mode (ATM); Video On Demand (VOD) network aspects

[5.4.10]: EURIM Briefing No19: The Regulation of Content on the Internet 1997. Juli

5.5. Network Services

Gabriella Süle dr., author

György Takács PhD, reviewer

Network services are provided to operators by carriers throughout network agreements. Occasionally network services are provided to subscribers using network parts. Considering economic aspects, the most important types of network services are leased line, interconnection, network access, and virtual private network. There are supplementary network services such as support for billing and Centrex service.

To provide network services signal transmission and processing are implemented between different products. There is a choice of standard interfaces to define the most important network services. Relevant inter-network interfaces provide for physical and logical levels of interconnection to maintain interoperability at network border. Communication services between terminal equipment used over interconnected networks are provided by interconnection traffic services. Network services for subscribers apply network-user interfaces that determine the characteristics of the communication services provided in the network to use by terminal equipment.

5.5.1. Leased Line Services

Leased line is composed of dedicated transmission circuits to use by a customer that provides for transparent transmission capacity between fixed network terminating points. It does not contain switching functions controlled by the user.

Leased line service consists in the provision of transmission capacity in the backbone network, or between customer site and backbone network of the carrier. Leased line services are often used by other operators for their network build out. In another usual application, a carrier connects the customer throughout a leased line rented from the other carrier, and provides subscriber services this way. In some cases mobile network operators use leased line as well to connect their base station

to the network. Business customer uses leased line services of different bandwidth to deploy its private network primarily for computer network or private branch exchange network applications. Access to the internet is provided to business customers by leased lines and can get internet connectivity of demanded speed.

Open standards for leased line services determine both internetwork interface and connection characteristics. These standards set the interface presentation, connection characteristics, and equipment interoperability requirements between network termination points. There are requirements for terminal equipment to be connected the leased line and related attachment conditions are standardised. Following tables show open standards for the most important analogue and digital leased line service offerings accepted in the European international practices and applied in Hungary as well.

	Technical Features	
Type of Leased Line	Interface Presentation	Connection Characteristics
64 kbit/s	ETS 300 288, ETS 300 288/A1	ETS 300 289
2 048 kbit/s – E1 unstructured	ETS 300 418	ETS 300 247, ETS 300 247/A1
2 048 kbit/s – E1 structured	ETS 300 418	ETS 300 419
34 368 kbit/s – E3	ETS 300 686	ETS 300 687
139 264 kbit/s – E4	ETS 300 686	ETS 300 688
N x 155 520 kbit/s – STM-N	ETS 300 299	Not available

Table 5.5.1 Digital Leased Lines

	Technical Features	
Type of Leased Line	Interface presentation	Connection Characteristics
Ordinary quality 2 wire	ETS 300 448	ETS 300 448
Ordinary quality 4 wire	ETS 300 451	ETS 300 451
Special quality 2 wire	ETS 300 449	ETS 300 449
Special quality 4 wire	ETS 300 452	ETS 300 452

Table 5.5.2 Analogue Voice Band Leased Lines

- Ordinary quality 2 wire leased line is previously provided in accordance with CCITT M.1040 Recommendation. The associated standard for terminal equipment is ETS 300 450. The attachment requirements for terminal equipment to be connected to these leased lines are described in ETSI Common Technical Regulation 15 (CTR 15).
- Ordinary quality 4 wire leased line is previously provided in accordance with CCITT M.1040 Recommendation. The associated standard for terminal equipment is ETS 300 453. The attachment requirements for terminal equipment to be connected to these leased lines are described in ETSI Common Technical Regulation 17 (CTR 17).

- Special quality 2 wire leased line is previously provided in accordance with CCITT M.1020 and M.1025 Recommendations. The associated standard for terminal equipment is ETS 300 450. The attachment requirements for terminal equipment to be connected to these leased lines are described in ETSI Common Technical Regulation 15 (CTR 15).
- Special quality 4 wire leased line is previously provided in accordance with CCITT M.1020 and M.1025 Recommendations. The associated standard for terminal equipment is ETS 300 453. The attachment requirements for terminal equipment to be connected to these leased lines are described in ETSI Common Technical Regulation 17 (CTR 17).
- For 64 kbit/s leased line the associated standard for terminal equipment is ETS 300 290 as amended by ETS 300 290/A1. The attachment requirements for terminal equipment to be connected to these leased lines are described in ETSI Common Technical Regulation 14 (CTR 14).
- For 2 048 kbit/s unstructured leased line the market denomination is E1. ETSI has developed the ETS 300 418 that replaces the ETS 300 246. The associated standard for terminal equipment is ETS 300 248 as amended by ETS 300 248/A1. The attachment requirements for terminal equipment to be connected to these leased lines are described in ETSI Common Technical Regulation 12 (CTR 12).
- For 2 048 kbit/s structured leased line the market denomination is E1. Previously provided in accordance with CCITT G.703, G.704 (excluding section 5) and G.706 (cyclic redundancy checking) instead of ETS 300 418 and previously provided in accordance with relevant CCITT G.800 series instead ETS 300 419. The associated standard for terminal equipment is ETS 300 420. The attachment requirements for terminal equipment to be connected to these leased lines are described in ETSI Common Technical Regulation 13 (CTR 13).
- For 34 368 kbit/s leased line the market denomination is E3. The associated standard for terminal equipment is ETS 300 689. The attachment requirements for terminal equipment to be connected to these leased lines are described in ETSI Common Technical Regulation 24 (CTR 24).
- For 139 264 kbit/s leased line the market denomination is E4. The associated standard for terminal equipment is ETS 300 690. The attachment requirements for terminal equipment to be connected to these leased lines are described in ETSI Common Technical Regulation 25 (CTR 25).
- For N x 155 520 kbit/s leased line the market denomination is STM-N in case of SDH VC based leased line bandwidth. The associated standard for terminal equipment is ETS 300 299. ETSI is currently revising ETS 300 299 and working on standards for this type of facility, they are preliminary specifications of prEN301 164 and prEN 3001 165.

In particular cases more than over one network and single operator provides the leased line service considering geographic extent of demanded service. Such case is the border crossing private network or a multisite company demanding leased line services from network operators over separated geographical areas. One stop

service provisioning is contracted by the customer and selected operator for the whole service demanded. For such service transactions leased line networks of the operators are physically interconnected. Service provision, fault clearence, customer care and billing are arranged by agreements among appropriate network operators.

5.5.2. Interconnection Services

Network interconnection is implemented by network operators' agreements to arrange provision and use of the network services over their networks. Implementing network interconnection, telecommunications network of the same or different network operators are physically and logically connected under defined conditions to allow the users of one network operator to communicate with users of the same or another network operator, or to access services provided by another organisation.

The largest network of extent is the internet and it works as global network of networks formed by a lot of interconnected networks. Interconnection of voice telephony and mobile radiotelephone networks is the most important for the public and from economic point of view. Networks of discrete telephony network operators, networks of mobile network operators, and furthermore voice telephony and mobile radiotelephone networks are interconnected. The reason is to allow making calls from any telephone network to any mobile network subscriber and vice versa, so subscribers of telephone network can be called from the mobile network. Network operators provide interconnection traffic services to each other in order to allow completing calls and using subscriber services between terminating points.

Network operators bill the calling party for originated traffic, and furthermore, they settle each other for measured interconnection traffic services. Characteristics of interconnected networks and telecommunication services offered to customers of different networks are of economic importance and determined by network architecture, traffic volume and distribution. Design of telecommunication services offered to customers over interconnected networks is based on cost calculation. The major part of operator costs comes from interconnection charges. In case of given traffic volume and distribution the structure and level of interconnection charge elements determine the economy of network architecture.

Interconnected network architecture can be economic in a case where the largest network of a geographical area is interconnected with all other smaller networks and customers of different networks communicate throughout the interconnection point of this largest network. In a different network environment and network deployment a different arrangement can be advantageous for network operators, this is the direct interconnection between networks and based on expected traffic distribution. Customers of different networks directly communicate each other throughout an interconnection point of the relevant two networks. Aim of regulation network interconnection and specially regulating the interconnection charges is to give measures for conditions of complete geographic and service coverage considering economic requirements as well.

Regulated interconnection traffic services are call origination or indirect access, call termination, transit, and number translation service. Overall technical description of these service provisions is as follows.

Call origination or indirect access service is used by carrier "A" when a subscriber of network "A" uses a selected subscriber service of this carrier and billed accordingly although call is originated in network "B". In the practice network "A" is an alternative network and network "B" a former monopoly operator's network. The subscriber of network "B" uses a short usually four digit carrier selection code to make call on network "A". Such call originates on network "B" and handed over to network "A" at the nearest interconnection point to the selected network.

Call termination is used by interconnected carrier when it hands over an originated call to the other carrier to reach the called party or service at network termination point on the other network and allow for communication.

Transit is used by an interconnected carrier to hand over a call from its interconnection point to reach another interconnection point. In this traffic case there is a need for a third network operator as well to terminate the call in its network. Specialised network operators provide transit services over long distance intercontinental paths.

Number translation service directs a call originated to a non-geographic number to a specific network terminating point where it is translated to geographic or mobile subscriber number for call completion purposes. To complete a call needs to

direct the call to a terminal equipment of specific geographic location or a mobile equipment of specific subscriber identification module.

Peering is non-regulated, initial interconnection form with simplified settlement where traffic exchange between networks is based on barter arrangement. This form is primarily used in equal traffic distribution and where the traffic termination cost is less than costs of traffic measurement and billing. Number of customers, or network computers and applications can be further considered for setting conditions of peering agreements. Today this form is used for internet services.

Open standards for inter-network interfaces applied for interconnection shown by following table.

Table Interconnection interfaces for PSTN telephony, ISDN, and GSM networks

Technical Characteristics	
Interface	SS7 Signalling System
ETS 300 356-1 to ETS 300 356-12	ISUP version 2
ETS 300 356-14 to ETS 300 356-19	
ETS 300 356-19/C1	
ETS 300 356-31 to ETS 300 356-36	
ETS 300 009-1, ETS 300 009-2	SSCP
ETS 300 008-1, ETS 300 008-2	MTP
ETS 300 646-1 to ETS 300 646-4	Application if ISUP version 2 for the ISDN-GSM signalling interface

ISUP is the user part of Signalling System No. 7 (SS7). SS7 provides common channel signalling for use in circuit switched networks: PSTN, ISDN and GSM. It has been designed first at an international boundary, but also appropriate for the interconnections of different operators' network in the same country. ISUP uses the layers 1 to 3 protocols (MTP) and may also use SCCP. International organisations adopt new versions of these standards for general applications to meet developing demands of network operators.

5.5.3. VPN Virtual Private Network Services

VPN virtual private network service is a network service for business customers, primarily for voice transmission and computer network applications. Nowadays it is used for combined voice and data applications as well.

Private voice network service is a VPN made of private telephone branch exchange network and specialised to specific traffic distribution of the customer therefore its usage is cost-effective. Cost savings come from network provision over software based networking equipment instead of using dedicated cable pairs or radio channels. The customer has a subscription agreement with the network operator for implementation and operation of the virtual private network, and for the outgoing voice calls. Calls within the VPN network are free to the customer.

For data communications specifically computer network applications VPN virtual private network is deployed over the infrastructure and make it secure for customer's business critical applications as required. Using this service the customer is allowed to connect new service nodes to or disconnect existing service nodes from its VPN virtual private network on demand. Furthermore, customer can set temporary or permanent connection points. These networking functions can be extended also to third party, for example, its vendor and customers as well. Growth of internet and global development of telecom operators' managed IP networks have leveraged new VPN network type so called IP VPN. IP-VPN network operates as coded IP tunnel over managed IP-based private networks and of course throughout the internet.

Data security is provided fro customer by four functional elements of the VPN virtual private network:

- Authentication is ensuring that the data originates at the source that it claims, usually using user name and password
- Access control is restricting unauthorised users from gaining admission to the virtual private network
- Confidentiality is preventing anyone from reading or copying data as it travel across the internet or IP-VPN
- Data integrity is ensuring that no one alter data as it travels across the internet or IP-VPN

Customers prefer buying IP-VPN private network for business communication solution when the cost savings reach during two-three years typically thirty-eighty per cent. Buying decision is mostly based on service level agreement, such as network availability ratio, IP packet latency, and packet loss ratio.

5.5.4. Network Access Services

Network access includes the notion of interconnection although they are distinct service types in network operators' business practice and in telecommunication regulation. Network access provision is physically and logically connecting a telecommunication network to another one to allow the service provider to use services for providing services to its customers. Network access service in general means the making available of network elements and associated facilities, under defined conditions, on either an exclusive or non-exclusive basis. This may involve the connection of equipment, access to physical infrastructure including buildings, ducts and masts. The service may mean access to relevant software systems including operational support systems, access to number translation or systems offering equivalent functionality. Network access service may include access to fixed and mobile networks, in particular for roaming, access to conditional access systems for digital television services, and access to virtual network services.

Special network access provided over voice telephony network means the making available the connection of equipment to specific network termination points for the purpose of providing telecommunications services. These access points are different from subscriber network termination points. Such network access point, for example, maybe at a tandem switch in the voice telephony network under defined conditions of the access agreement. In this case IP-based services are made available for telephone subscribers on market demand and based on a specific service agreement.

Using transport facilities of the voice telephony network a network access service may be provided to use the transport and control capabilities of an IP-based network . For example, this is the case of remote access to the internet or an IP-WAN by using dial up access. Holding time of calls to the internet is longer then voice calls so the high growth internet traffic can overload tandem exchanges and cause damage of quality or even often call drops. Consequently, it may be necessary to change the existing architecture of voice telephony network, furthermore to divert a part of internet traffic from tandem network.

Another network access service provides for interoperability of voice telephony and managed IP-based private network (or the internet network) at transport and

control layer. Such converged network service is IP voice between telephone equipment, so called voice over IP (VoIP).

Open standards for network access interfaces are not available yet. Preliminary technical specifications are available for service provider access to network functions in third generation mobile networks, VPN virtual private networks, broadband bearer services, specific multimedia and internet services.

5.5.5. Network Service for Billing Support

Because of telecommunication service liberalisation there are many services available over networks demanding billing and customer care functionalities as well. Network operators of interconnected networks need co-operation in tariffing and charging calling party. Further processes of billing, like printing and issuing the account, furthermore collection and billing claims are the operator's job.

Network service for billing support may include information provision of call records from mediation system of the network operator. Another network service is charging and tariffing either for interconnection settlement between network operators or for subscriber charging using list prices. Subscriber price plan, discounts and promotion are taken into consideration by the carrier who has concluded subscriber contract.

5.5.6. Centrex Service

This service is named after private in-house telephone system services provided by telephone carrier by means of facilities of local telephony exchange. Traditional Centrex service is used by specific, consequently costly terminal equipment and all users need it similarly to programmed telephone equipment of private in-house telephone system. Another consequence of the service structure and operation is all calls set up and directed over main lines of the voice telephony operator and two main lines is used for in-house calls. These conditions prove economic only in case of specific traffic pattern. Particular customers, like university campuses, make practically long distance and international calls from in-house telephone system, and not in-house calls, so Centrex can be cost effective for them.

New development opportunity for implementing Centrex is IP-Centrex. Similarly to other IP-based telecommunications, due to service quality and security requirements, it can be provided not only over the internet but also over managed IP-based network of the operator. On the contrary to the port-based analogue and digital voice telephony services, software-based facilities are dominant in the unit cost of Centrex service provisioning, and hardware costs attached to the number of user are much less.

Service is implemented by network operator to meet service quality characteristics required by subscriber. Traditional voice quality and dialog characteristics can be set by gateway facilities and IP bandwidth. Both calls to the voice telephony network and between IP networks call forward, call waiting, voice mail, call center and other value added features are available for users. As system design, IP-Centrex can be considered as a specific IP-VPN specialised to voice service provisioning only. IP-based global networks provide IP Centrex services for employees, clients and suppliers of the organisation using this service as they would use it at the headquarters. This is a kind of telecommunications mobility and can be limited by user organisation by setting service security elements accordingly.

Abbreviations:

ITU-T

SDH

References

Further readings are offered for readers who interested in details:

[5.5.1] Darryl P. Black: Building Switched Networks; Addison-Wesley, 1999

[5.5.2] Igor Faynberg - Lawrence Gabuzda - Hui-Lan Lu: Converged Networks and Services; Wiley Computer Publishing, 2000

5.6. Terminal equipment, services implemented in terminal equipment

Lajos Pomázi, author

Erika Madarász, reviewer

The subscriber telephone sets of wired telephone networks can be considered as mature and settled among terminal equipment. Continuous changing, modification and renewal characterise the terminal equipment of mobile telephone networks that are integrant part of the system and can be hardly handled independently from the other parts of the system. Therefore this chapter contains the relevant elements that are essential on the fields of services offered by terminal equipment and subscriber sets and shows their application examples via traditional telephone sets.

5.6.1. Subscriber interface

Different kinds of terminal equipment are connected to the exchange via subscriber interface. The main circuitry functions of subscriber interfaces can be summarised by “BORSCHT”, where the meaning of these letters is the following:

B: Battery supply, Battery Feed; Supply of the subscriber loop

O: Overvoltage protection;

R: Ringing; Ringing signal applied to subscriber line

S: Supervision, Signalling; Checking/Monitoring of the status of subscriber loop

C: Coding; A/D, D/A conversation (coding, decoding)

H: Hybrid, 2-wire to 4-wire/4-wire to 2-wire conversation

T: Testing; Testing of the subscriber interface

Feeding of the subscriber loop: Feeding bridge supplies terminal equipment with direct current via subscriber line. The feeding impedance is sufficiently high to

ensure only a small loss, and thus feeding bridge has little or no effect on transmission.

Because of traditional reasons the DC parameters of conventional resistance feeding bridges can differ for each country [5.6.1], moreover for the different types of telephone exchanges established in the same country. The most frequent DC parameters of telephone exchanges in Hungary are 48V, 2x250 ohms, 48V, 2x300 ohms and 56 V, 2x220ohms. Depending on the parameters of the terminal equipment the maximal line current can reach 75 mA – 100 mA at very short subscriber line length, therefore some Watts of power can also dissipate on the feeding bridge.

In order to reduce the energy requirements, optimised constant current feeding bridges have been introduced. The maximal feeding current is limited to a lower value (i.e. 22 mA, 30 mA) that can be maintained in a given range of the subscriber loop resistance. When the loop resistance is out of this range, it works as a conventional resistance feeding bridge.

Due to the line current attenuation, depending on the type of the feeding bridges the maximal loop resistance of the subscriber line is in the range of 1600 ohms and 2200 ohms.

Overvoltage protection: Its task is to protect the sensitive electronic circuits against the damaging effect of overvoltages coming from the subscriber line (i.e. induction of alternative voltage derived from power-supply system, surges due to lightning, electrostatic discharges). The protection has to come into action within some nanosecond when the voltage reaches the specified level. Protection against overvoltages shall fulfil the relevant ETSI standard [5.6.2] and ITU-T Recommendation [5.6.3].

Ringing: During incoming calls the ringing circuit gives ringing signal to the subscriber line. The ringing signal can be characterised with its voltage (typical value: 75 V – 90 V), frequency (typical value: 25 Hz) and cadence (on: 1250 ms, off: 3750 ms). After the incoming call has been answered, the ringing has to be tripped within some ten milliseconds.

Coding - decoding: The coding and the decoding circuits are placed in the 4-wire path. The encoding circuit established in sending direction takes sample in every

125 μs from the analogue signal gone through a low-pass filter and digitises this sample into 8-bit PCM code using a special encoding technique in accordance with A-law [5.6.4]. The decoding circuit established in receiving direction converts the 8-bit PCM encoded digital words into analogue signal, that arrives at the hybrid circuit via a low-pass filter.

2-wire to 4-wire conversation: Primary use of hybrid circuit to convert between 2-wire and 4 wire operation in concatenated section of a telecommunication circuit. The appropriate impedance matching between the balance network and the terminated impedance of 2-wire port reduces the signal level returned via hybrid, the echo. The requirements for 2-wire analogue interface are given in ITU-T Recommendation Q.552 [5.6.5]

Testing of the subscriber line: From the point of view of operation and maintenance it is desirable that the different circuits of subscriber interface and the subscriber line can be tested independently of each other.

Requirements for the interworking and equipment connected to an analogue subscriber interface can be found in [5.6.1].

5.6.2. Handset requirements

Handset is a standard telephone component, that includes a telephone microphone and a telephone receiver. User holds it in hand close to the head during telephone conversation.

The profile and the dimension of the handset must be such that

- the earpiece shall fit comfortably to the user's ear and the mouthpiece must be close in front of the user's lip;
- the handgrip shall be convenient to hold it and shall allow sufficient room for fingers to wrap and clearance for cheek.

The shape and the dimension of the handset influence the sending and receiving levels of the telephone sets. Based on the information gained by mass measuring the ITU-T recommended the size and the shape of handsets according to Figure 5.6.1. [5.6.6] [5.6.7]. The investigations show that, for convenience in use, the mouthpiece should be 10-12 mm far from the circle X enclosing the centre of the lip of 80 % of the subjects tested (over 4000). Mouthpiece should touch the circle Y in

such a way that the tangent line should be greater than 30° to the speech direction. A handset conforming to Figure 5.6.1 is acceptable to more than 90% of users.

The majority of manufacturers takes this recommendation into consideration to design handsets harmonising with telephones. However in some special cases their function and/or appearance are the major consideration such as mobile phone or telephones shaped as banana, shoe, hamburger, etc.

The casework should be robust and rigid and not susceptible to flexing and creaking when handled. Split lines (the join between the sections of the casework) within the handset moulding do not cause sharp edges or skin pinching.

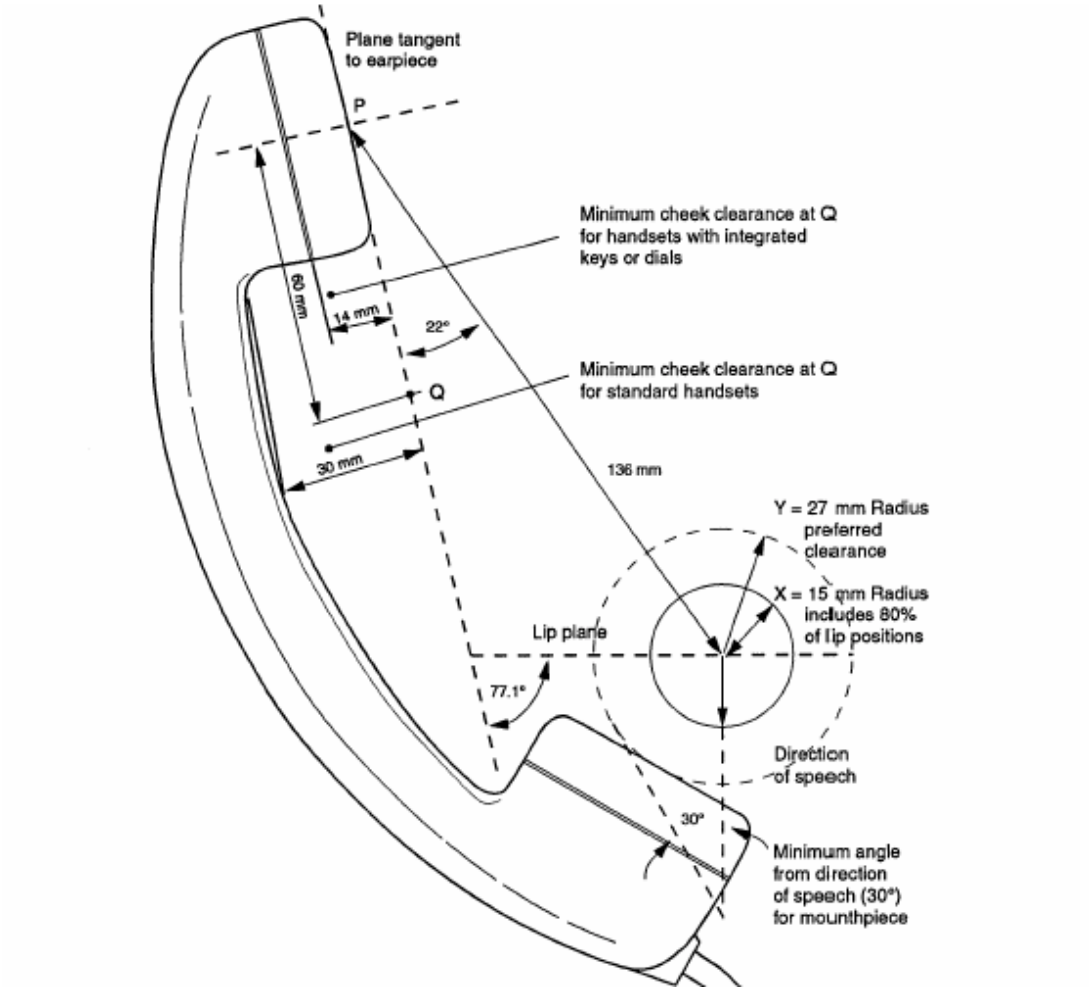


Figure 5.6.1. Preferred handset profile dimensions according to ITU-T Recommendation P.35

5.6.3. Hands-free telephone

Hands-free telephone is a telephone set using a loudspeaker associated with an amplifier as a telephone receiver and a built-in microphone in the telephone house associated with an amplifier as a telephone microphone and which makes it possible to make a telephone call without having to hold the handset during the call. When the handset is picked up during hands-free mode, the telephone has to switch off the hands-free operation mode automatically and has to switch over to handset mode.

According to the subjective expectation

- hands-free telephones shall have adequate sending and receiving sensitivity;
- singing through acoustic feedback between microphone and loudspeaker has to be avoided;
- the degradation of the speech quality due to the voice switching shall be negligible.

Similarly to handset telephones the sending and the receiving sensitivity of hands-free telephones can be expressed in terms of Loudness Rating. Because distributing LR values between the telephone sets and the network is within the competence of the national transmission plan existing in most countries, ITU-T does not issue an international recommendation stating LR values of telephone set alone – whether these are handset or hands-free telephone. However it is possible to recommend the sending and receiving sensitivity for hands-free telephone relative to the standard handset telephone set used nationally, having regard to the physical difference between handset and hands-free telephones and the effects of different user behaviour [5.6.8] [5.6.9].

The sending loudness rating (SLR) of a hands-free telephone should be about 5 dB higher than the SLR of a handset telephone.[5.6.8] The difference of 5 dB has several components:

- the average speaking level is about 3 dB higher for hands-free telephone than for handset telephone;
- the output level from a handset telephone in conversational use, which is about 1-2 dB lower than what is obtained in the speaking position specified for loudness rating measurements;
- the different frequency responses of the microphones cause significantly smaller differences than the components mentioned above.

It should not be possible for the user to adjust the sending sensitivity [5.6.8].

For the objective evaluation of the sending loudness of hands-free telephone

- measure the sending frequency sensitivity curve of the hands-free telephone placed in the physical test arrangement described in §6 of ITU-T Recommendation P.34, then
- calculate the sending loudness rating according to the method described in ITU-T Recommendation P.79 [5.6.10].

According to the ITU-T Recommendation P.34 the receiving sensitivity of a hands-free telephone without automatic gain control should be adjustable within a range of 15 to 30 dB. This range should include the value of the receiving loudness rating (RLR) which

- is equal to the nominal RLR of the corresponding handset telephone, and
- is 10 dB better than the nominal RLR of the corresponding handset telephone.

10 dB lower receiving LR than the nominal RLR of the standard handset telephone is necessary to compensate the impact of up to medium high background noise level. The preferred listening level depends on the background noise level. According to the subjective opinions test in case of background noise higher listening level is required for hands-free telephones than for handset telephones in order to get the preferred listening level. [5.6.8] [5.6.9]

The maximal receiving amplification has to be limited in order to avoid the overhearing of other telephone conversations due to crosstalk.

For the objective evaluation of the listening loudness of hands-free telephone

- measure the receiving frequency sensitivity curve of the hands-free telephone placed in the physical test arrangement described in §6 of ITU-T Recommendation P.34,
- calculate the receiving loudness rating according to the method described in ITU-T Recommendation P.79 [5.6.10], and
- subtract 14 dB from the computed loudness rating. This correction factor contains the appropriate correction for the difference between one-ear and two-ears listening (12 dB) and the difference of about 2 dB caused by loudspeaker listening versus binaural earphone listening, due to the diffraction effect of the listener head [5.6.9].

Most hands-free telephones contain voice-switched circuit, whose main purpose is to avoid the singing through the acoustic feedback that can be arisen between microphone and loudspeaker. Such circuits insert a loss in either the sending or receiving direction. Switching from one direction to the other one occurs

- when signal above a given threshold is applied from the opposite direction, or
- when the control circuit, taken into account the relative levels and the nature of the signals in both direction, allows the switching.

By a suitable choice of the thresholds and the switching times, the degradation of speech quality (i.e. clipping effects and loss of initial or final consonants in the transmitted speech) that is introduced by voice switching can be made negligible. [5.6.9]

A possible circuit description can be found in [5.6.11] gives the details

- how the gain of the sending and receiving amplifiers is controlled depending on the signal and the noise levels in both directions, in order to avoid the singing;
- how the switching times are set in order not to worsen the subjective opinion of speech quality by voice switching.

5.6.4. Keypad/pushbutton requirement

Telephone keypads and keyboards are the basic means for providing access to telecommunication services in the most telecommunications equipment. Keypads consist of keys arranged according to certain principle. Keys are labelled with digits and/or letters and maybe symbols indicating their function.

According to ITU-T Recommendation E.161 [5.6.12] the recommended minimal keypad of terminal equipment includes 12 keys, the numeric keys 0 - 9, star (∗) and square (#). ITU-T gives some possible versions for the keypad arrangement, but the layout can be seen in Figure 5.6.2. is considered as the preferred solution. ETSI takes a strong stand for the standard 4x3 keypad array. [5.6.7]

Unsighted navigation on the keypad is helped when one of the buttons can be discernible from the others by touching. The preferred form of the tactile identifier in the middle of the key “5” is a distinct raised round dot. The recommended dimension of the raised round dot is 0.6 mm (±0.2 mm) high and with a diameter of 1.5 mm (± 0.2 mm) [5.6.13].

In general it is advantageous when the keys can not be pressed below the surface of the keypad. Key should have a minimum surface of 113 mm^2 with a minimum the dimension on any direction of 12 mm. The key travel should be in the

range of 1 mm and 8 mm. The pressure necessary to activate the keys should be 0.25 N - 1,5 N.

It is advantageous for users if suitable key feedback is given in not only tactile but also acoustic form, i.e. when the key is activated, a noticeable click can be felt and heard from it. The preferred pith (distance from key centre to key centre) on the keypad is 19 mm ± 1 mm. If this parameter is less than 15 mm then it can lead to significant increases in miskeying [5.6.7].

It can be helpful in memorising telephone numbers or supplementary service codes if also letters are assigned to the numeric keys on the keypad. ETSI defined the assignment of alphabetic letters to digits in a standard telephone keypad array are shown in Figure 5.6.2. [5.6.14] The legibility of the numeric digits should not be impaired by alphabetic letters. If the keypad is used for entering of letters as well (i.e. processing of names in case of electronic telephone book) then visual feedback is strongly advisable. The user should never be expected to key data “blind”.

Depending on the services of terminals the standard 4x3 keypad can be supplemented by additional function keys. It is advisable to separate the additional function keys from the keys of standard keypad. Function keys should be labelled with the full function name in the national or preferred language and/or with well-known symbols. ITU-T gives the symbols for the most frequently used services and functions in [5.6.15]

In order to reduce the numbers of function keys in complex devices, softkeys can be applied. They are physical or software displayed keys whose functions are

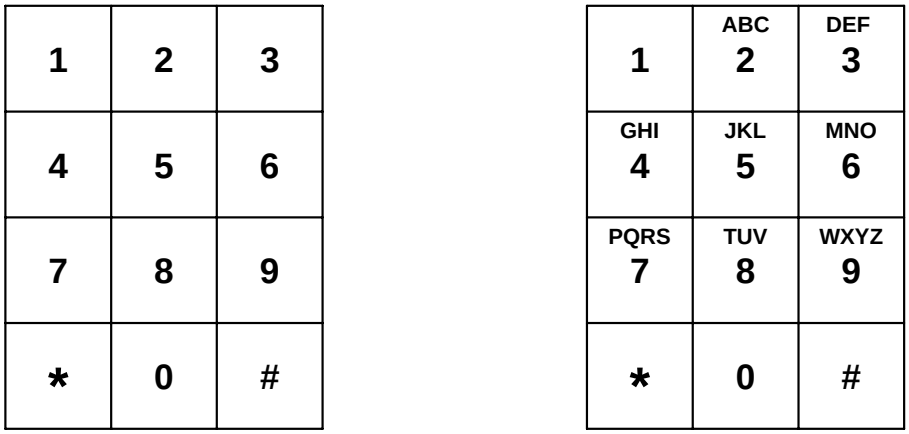


Figure 5.6.2. Numeric layout and layout for alpha characters on telephone keypad

determined by the application software currently running on the terminal. Soft keys labels shall be as physically close to the key as possible. These labels should be explanatory, as far as possible avoiding abbreviations.

5.6.5. Display requirements

The visual display is an essential element in the design of user interface and the most flexible medium ranging from simple light signals used to give status and warning information, through alphanumerical character displays to extensive text and live pictures shown on large CRT (Cathode-Ray Tube) and LCD (Liquid Crystal Display) graphic displays [5.6.7]. Visual displays are used to provide feedback to control input and data entry, to give prompts and system messages and to show text, graphics and pictures.

During the design of terminal equipment the choosing of displays should be influenced by amount and nature of information, size and legibility of displayed characters and viewing distance. Taking account of the degree of importance of displayed data as well, the recommended minimum character height in function of the viewing distance is given in [5.6.16].

Character displays have typically from 1 to 8 lines by max. 24 characters per line. LCD displays are mostly selected, that are used in a wide variety of products for displaying of numbers, alphanumeric characters and symbols indicating set-up and/or status information (i.e. dialling mode, battery charge, field strength, message arrived) of terminals. The simplest 7 or 9 segment displays are acceptable for display of numeric information. Dot –matrix is necessary where alphanumeric information is required. As a minimum, a matrix of 7x5 shall be used for capital letters. Dot matrix of 9x7 with the addition of four rows to accommodate line spacing, lower case ascenders and descenders and accents is more advantageous (i.e. “g”).

Graphical displays provide a dot-matrix or similar technology to display static or dynamic images, including text, graphics, symbols and pictures.

In favour of legibility the character/background contrast ratio should be 3:1 minimum, 10:1 recommended [5.6.16]. Light weight, extra bold or condensed typefaces should be avoided. Adjustability of the contrast depending on the

environmental temperature and light is advantageous feature. Backlighting function for LCD used in low ambient light should be provided.

Icons (possibly standardised) should be used whenever possible for identifying common functions or objects on the display. Good icons are easier to understand and remember by most users, and can be recognised also by people with reading difficulties.

5.6.6. Requirement for automated dialogue services

A user interface or a Man-Machine interface is the bi-directional interface through which a user communicates with a telecommunication terminal or via a telecommunication terminal to a telecommunication service provider. ETS 300 738 describes the requirements of the minimum user interface for use to gain access to and control of supplementary services within public telecommunication networks [5.6.17].

A variety of supplementary services (i.e. voice mail, telebanking, information retrieval, shopping) are available that are accessible via public telecommunication network, that users can control with DTMF signalling from the standard telephone keypad and where the system informs the user about the system status and the result of control functions with stored voice announcements. The ETSI Technical Report of ETR 329 contains the guidelines for procedures and announcements in the Stored Voice Services [5.6.18]

Depending on the format in which a service requires the data input from the user, the dialogue between the user and the system can be

- command dialogue
- interactive dialogue

In a command dialogue the user control a service by entering a complete sequence of digits and symbols that includes all the necessary information without any prompt or interaction with the service. The user shall know the syntactic of command, as he/she is informed about the result only after the entering of the last character whether the command is accepted or refused by the system.

In an interactive dialogue no prior knowledge of the service or its command syntax is required for the user to use the service, as the interactive dialogue guides

him/her through all the steps. Each time the user reaches a new service state or stays idle, the service prompts for the next user actions.

The choice between two dialogue types depends on the complexity of the service and the tasks the user expecting to carry out. If the service only offers a single option and requires only some input information from the user, a command dialogue is appropriate. When a variety of options is available (i.e. banking service), the system requires a lot of input information from the user (i.e. card number, PIN code) or users' guide among menus is recommended, an interactive dialogue is preferred.

From the point of view of users the stored voice services should fulfil the following requirements:

- the user should feel that it has been satisfying and have confidence in the outcome;
- private data cannot be accessed by unauthorised person;
- services should be easy to learn and handle, the users should not be expected to learn lengthy, unstructured command strings;
- user shall get enough information about the actual state of service, its control feasibility and the result of the previous control action (i.e. instructions, prompts, feedback, error cue, etc.);
- error handling and correction shall be possible without detriment to the service and without feeling it necessary to hang up and start again;
- the most frequently used option should be accessible in the higher levels reducing the time spent in the system.

General design principles of stored voice services are the following:

1. Design of the structure of a dialogue

- collect default information from all available sources. Methods, that minimise the number of manual input operations to be performed, are preferred. (i.e. Calling line identification);
- user identification and/or authorisation;
- facility for language selection;
- choice of the structure of menus depending on the complexity of service;
- building-up of menus.

2. Design of user control over a dialogue

- use of the principle of "Indicate - Control – Indicate". User shall be informed about the actual state of service before control action, he/she shall knows the

control commands available in that state and he/she has to get feedback indication after his/her control action about the changed state of the system;

- “cut-through” announcement. The system should provide an opportunity for users to enter the command as well while an announcement is being played. When a valid command is entered during an interruptible announcement, the announcement should be interrupted and the system should go to the same state as if the command had been entered at the end of the announcement. The “cut-through” technique can not be used for the announcements, which inform the user of unexpected occurrences. (i.e. access to the mailbox N is not possible, since mailbox N does not exist);
- “dial-ahead” function. Dial-ahead allows the experienced user to enter ahead of time several commands in a sequence and the system should bring the user directly to the corresponding state;
- choice of system response time. The delay from the receipt of a user input to the audible start of the corresponding system response should be in the range of 500 ms and 1 s;
- handling of illegal control code. (i.e. the system responds as if no key had been pressed at all or error announcement);
- error management;
- choice of time-outs (from 2 s to 8 s) and the handling method of time-outs (i.e. repeat of the announcement of the actual service state);
- asking for confirmation when an action is going to change the system in an irreversible way (i.e. stored data will be erased, execution of transaction);
- design of data entry;
- design of general function (help, repeat, pause, go back to the main menu, etc.).

3. Design and wording of announcements used in stored voice services

- the words used for all announcements should be quickly and easily understood by all possible users of the service;
- (technical) jargon, words with double/unclear meaning and homonyms (words which sound similar) should be avoided;
- when giving instructions to the user, the imperative form of the verb should be used;
- important information should be given early in an announcement. In a menu item, the anticipated result should be indicated before the command. When presenting a list of related items, “or” should be used before the last item in a list;
- sequences of actions presented in announcements to the user should follow the order in which the user is expected to carry them out;
- feedback announcements should be short and explicit;
- numbers should be announced according to its natural spoken form;

- tones used in stored voice services should fulfil the relevant ITU-T Recommendations [5.6.19] [5.6.20].

4. The characteristics of the speaker recording the voice announcements should:

- speak clearly and rhythmical, articulate well,;
- be assertive and agreeable, be practised at microphone speech;
- be female voice preferably.

5. At recording of announcements the following aspects should be observed:

- ensure sufficient technical background (adequate bandwidth, minimal frequency distortion and background noise);
- ensure the speech speed (words per minute) is acceptable for high levels of comprehension;
- ensure the speed, volume and intonation of separate blocks are such that they sound “natural” when put together;
- avoid combining artificial and recorded speech in the same sentence, if it can not be avoided insert a short pause (1 to 2 s) between the two forms of speech;
- when combining recorded messages and recorded announcements (i.e. voice mail), ensure there is a short pause (1 to 2 s) between the two forms to emphasise the different sources of the material;
- preferred listening level for announcements should be -10 dBPa ± 5dB[5.6.20].

5.6.7. Device intelligence

Users can chose from a wide variety of equipment taking account of design, services and price. Sorting by their functions, the following services can be highlighted:

Services related to dialling:

- Flash;
- pause inserting between digits;
- last number redial;
- automatic redial of the last calling number by pushing of a dedicated button (auto redial);
- redial of a number selected from the last “n” telephone numbers dialled;
- call from memory;
- voice controlled dialling.

Storing of calling numbers

- storing of calling numbers to short codes;
- storing of calling numbers to direct memories;
- storing of calling numbers with names (telephone book function);
- noting of telephone number during conversation.

Functions related to call restriction

- restriction of calling number(s) or calling direction(s) with key or PIN code;
- calling a pre-programmed calling number by pushing of any button;
- locking of keypad.

Services related to speech

- microphone muting;
- volume control of receiving loudness;
- loud-speaking of receiving direction;
- hands-free telephone;
- holding of the telephone conversation;
- inform the other party about holding of the line with music.

Services related to call indicator

- setting of the loudness, the tonality and the cadence of incoming call indicator;
- visual display of incoming call.

Services related to display (calling line identification is mentioned in other section)

- visual display of the active status of telephone;
- display of called number;
- display of the call duration;
- display of the tariff charge for actual call;
- display of total tariff charge for calls have been made;
- display of date and time;
- language selection option;
- display of the status of telephone set (i.e. accumulator charging level, field intensity, dialling mode);
- backlighting of display.

Data protection with PIN code

Services related to answering machines

- recording or “announcement only” mode;
- recording and checking the outgoing message, measuring of recording time;
- setting the number of rings before activation of the answering machine;
- setting the recorded message length with or without time limit;
- monitoring of incoming messages;
- stamping of incoming messages with day and time during recording and its indication with artifical speech at listening of incoming messages;
- displaying the number of new messages and the recording time remaining;
- listening to the messages, stop and restart of the play-back, repeated play-back, skip to previous message, skip to the following message, storing/deleting message(s), deleting all the messages;
- remote retrieval with PIN code;
- recording the current telephone conversation.

Some special services of DECT cordless phones

- registration of max. 6 handsets to base unit;
- possibility of external and internal (intercom) conversation in the system in the same time;
- transferring a call from one handset to another;
- conference call;
- calling party identification based on stored voice announcement;
- baby-sitter function (unrepresentative service). When the background noise level overpasses the threshold adjusted, the handset calls automatically the destination handset within in the system. the destination handset answers this call automatically.

Some typical services based on calling line identification presentation (CLIP) [5.6.21][5.6.22]

- calling party identification based on the caller’s telephone number;
- calling party identification based on the caller’s name attached to the phone number. Precondition of this service is that the telephone number together with the name have to already be stored in the telephone-book of the device;
- getting information about the cause of missing telephone numbers (caller prevents the displaying of his name and number or caller calls from an area where this service is not offered);
- calling list with the entries of the last “n” unanswered call. A calling list entry contains the call number, maybe also the number of calls from the relevant call number, date and time;

- calling list with the entries of the last “y” answered call. A calling list entry contains the call number, maybe also the number of calls from the relevant call number, date and time;
- differentiation of callers or categories by different specific ringing tones (i.e. VIP call);
- connecting caller-ID device to computer. Caller-ID device sends the received data to the computer via serial port. Based on this information the computer can display also the caller’s detailed data.

5.6.8. Special needs of elderly and disabled people in relation to telecommunications

Telecommunication services have to be made accessible to as many people as possible including people with innate or acquired deficiency. Deficiency can be caused by physical, mental or sensory impairment, physical conditions or distemper [5.6.23]. New products and services should meet also the special needs of elderly or disabled people not to exclude them from taking part in society on equal terms with the rest of the society.

ETSI [5.6.24][5.6.25][5.6.26] propose to

- identify some of the main factors that can be inhibit the access to and use of telecommunication services for people with special needs;
- work out recommendations that take into consideration these special needs in order to make basic telecommunication services accessible to as many people as possible.

This chapter can offer only a brief survey of this very extensive field.

Visually impaired people can be divided into three category:

- people who are blind or seriously visually impaired;
- people who are partially sighted and who cannot read;
- people with reduced vision but who read with aid.

With regard to telecommunication, visually impaired people especially belonging to the first two group may have great difficulties in locating a terminal in unfamiliar surroundings or handling of unknown equipment.

Locating of public payphones can be facilitated by showing their position on direction boards or floor plans or by specially patterned pavement or floor surfacing which can be easily perceived by blind people. Insertion slot(s) for coins/cards must

be fashioned in a tactile manner for easy identification. Card based means of payment should be tactilely marked to identify easily the position and the direction of the card.

Beyond the visual display messages there should be some form of acoustic feedback that can also i.e. indicate the acceptance of means of payment inserted or give warning tone for the user in time when coin(s)/card in the payphone will be expired in a short time and new coin/card has to be inserted to maintain the existing conversation.

Preliminary condition of the faultless data entry is that the digits on the dial are laid out in standardised way [5.6.7] [5.6.12] [5.6.14], a tactile raised “dot” on the touching surface of the centre key “5” should be provided to help unsighted navigation of the keypad [5.6.13]. Sufficient space should be between the keys. The height of the characters on keys should preferably be over 7.5 mm. Clear visual contrast should be as high as possible between the keys and the legend of the keys. [5.6.26]

Between 7% and 10% of all males in the normal population are colour blind or have deficient colour vision. Many visually disabled people also have reduced or no colour vision. It is, therefore, recommended that colour should never be used alone to indicate vital functions and messages, but always in addition to other modes of information. If colours are used to identify or separate keys and information on displays, the colours should be chosen so that they are easy to separate into distinct grey-tones.

Approximately 10% of Hungary’s population have some short of degree of hearing impairment. Around 300,000 of them are seriously hard of hearing, besides there are 60,000 people who are deaf [5.6.23] [5.6.27]. People with impaired hearing can be divided into two main categories according to the seriousness of the hearing loss in relation to telephone communication:

- people who are hard of hearing;
- people who are profoundly deaf.

Depending of the seriousness of hearing impairment various solutions and devices can be needed providing access to telecommunication services for a large range of disabled people as well.

Even a moderate hearing loss can make it difficult to hear a ringing telephone. Amplification of the ringing signal, combining remote bells, and alerting lights on the telephone or remote from it can be useful.

The communication ability of deaf people depends on several factors [5.6.23]. These factors can be the time of the onset of deafness, the hearing has been lost before or after they have learnt to speak (pre-lingually or post lingually deaf), faculty of speech, degree of retained hearing, ability of lip-reading, personality, education, etc.

Deaf people with no or unintelligible speech can use the written communication. After set up of a modem connection between the parties, they can change messages via text telephones provided with alphanumerical keypad and display or computers operating with applicable communication software. Facsimile, SMS and electronic mail can also be used for providing information, however the information quantity, that can be considered as a logical unit, has to be sent in the scope of new address. Video telephone for manual sign language communication and lip reading support may soon offer good help.

The precondition of communication in the cases mentioned above is to have the same or compatible technical platform at both parties. By the help of special operator service it is available for the people with hearing or speech impairments to communicate with anybody via operator assistance, independently of the fact, that who has originated the call [5.6.28].

The communication of the people who are hard of hearing via telephone can be helped by telephone sets providing for

- inductive coupling to hearing aids;
- adjustable, additional acoustical amplification.

The recent hearing aids are equipped with induction pick up coils, wherewith can detect the alternating magnetic field with special characteristics generated by the magnetic induction systems incorporated in telephone handset. Hearing aids amplifies this induced signal and converts it into acoustical signal. In order to obtain proper interworking between telephone receivers and hearing aids via inductive coupling, the alternating magnetic field strength generated around the telephone earphone capsule shall fulfil the relevant ITU-T Recommendation and ETSI standard

[5.6.29] [5.6.30]. This magnetic field strength shall be high enough to produce an acceptable signal to noise ratio but not so high as to cause overloading of the hearing aid.

Telephone sets that provide additional receiving amplification can be suitable for the people who are hard of hearing. With the maximum additional receiving amplification selected, the receiving sensitivity should be at least 20 dB higher compared with the receiving sensitivity of normal telephone sets [5.6.29]. As hearing impaired people do not necessarily have elevated threshold of loudness discomfort, some form of output limitation is required. Automatic Gain Control may provide a better means of limitation than peak clipping. The acoustical howling can be avoided by the methods used at speakerphones.

Where the telephone is used by persons with a range of hearing acuity, it is preferable to reset the gain of receiving amplification to the nominal level automatically when the handset is replaced. Where the telephone is to be used mainly by a hearing impaired person, it is advantageous to maintain the receive amplification when the handset is replaced.

The main problem for mobility impaired people is actually getting to the telephone so that they can use it. Physical access to telephones in the homes or the workplaces of mobility impaired people can usually arranged in such a way that it is no great problem to reach the telephone, nevertheless the answering of incoming calls may need some time. The cordless telephone or even the mobile telephone is of great help to these people.

Public telephones and telephone boots must be designed for full accessibility to people in wheelchairs and people who can only walk with help. There must be no steps which prevent the access to the public phones. Doors must be easy to open and close by the people sitting in the wheelchair. The width of the boot and its door must allow entry of a standard wheelchair. The telephone must be placed at a height, where all controls (handset, dialling and function key, coin/card insertion slot, coin refund slot, etc.) can be operated from wheelchair and display messages of the terminal can also be read from this position.

The payphone location described above can be suitable for children and the people of shorter than the average stature. It is obvious that the optimal payphone

location for the people who are much taller than the average is different from the previous one. If several payphones are installed close to each other, they should be placed in such a way that people who are shorter or taller than the average can also find a suitable terminal.

References

[5.6.1] ETSI EN 300 001 (1998-10): Attachments to Public Switched Telephone Network (PSTN); General technical requirements for equipment connected to an analogue subscriber interface in the PSTN

[5.6.2] ETSI EN 300 386 (2000-03): Electromagnetic compatibility and Radio spectrum Matters (ERM); Telecommunication network equipment; ElectroMagnetic Compatibility (EMC) requirements

[5.6.3] ITU-T Recommendation K.20; Resistibility of telecommunication equipment installed in a telecommunication centre to overvoltage and overcurrents (02/2000)

[5.6.4] ITU-T Recommendation G.711; Pulse Code Modulation (PCM) of voice frequencies (Blue Book)

[5.6.5] ITU-T Recommendation Q.552; Transmission characteristics at 2-wire analogue interfaces of digital exchanges (11/96)

[5.6.6] ITU-T Recommendation P.35; Handset telephones (Blue Book)

[5.6.7] ETSI ETR 116 (1994): Human Factors (HF); Human factors guidelines for ISDN Terminal equipment design

[5.6.8] ITU-T Recommendation P.34; Transmission characteristics of hands-free telephones (03/93)

[5.6.9] CCITT Handbook on telephonometry, Geneva, 1992

[5.6.10] ITU-T Recommendation P.79; Calculation of loudness ratings for telephone sets (09/99)

[5.6.11] Philips Semiconductors, TEA 1098 Speech and handsfree IC (2000 Sep 26)

(http://www.semiconductors.philips.com/comms/products)

[5.6.12] ITU-T Recommendation E.161; Arrangement of digits, letters and symbols on telephones and other devices that can be used for gaining access to telephone network (05/95)

[5.6.13] ETSI ES 201 381 V1.1.1 (1998-12): Human Factors (HF); Telecommunication keypads and keyboards; Tactile identifiers

[5.6.14] ETSI ETS 300 640 (August 1996): Human Factors (HF); Assignment of alphabetic letters to digits on standard telephone keypad arrays

[5.6.15] ITU-T Recommendation E.121; Pictograms, symbols and icons to assist users of the telephone service (07/96)

[5.6.16] ECMA-136 (1989), Ergonomics - Requirements for non CRT Visual Display Units

[5.6.17] ETSI ETS 300 738, Human Factors (HF); Minimum Man-Machine Interface (MMI) to public network based supplementary services

[5.6.18] ETSI ETR 329, Human Factors (HF); Guidelines for procedures and announcements in Stored Voice Services (SVS) and Universal Personal Telecommunication (UPT)

[5.6.19] ITU-T Recommendation E.182; Application of tones and recorded announcement in telephone services (03/98)

[5.6.20] ITU-T Recommendation E.183; Guiding principles for telephone announcement (03/98)

[5.6.21] ETSI TR 101 292, Public Switched Telephone Network; Protocol over the local loop for display and related services; Proposed enhancements and maintenance of existing standards (1999-09)

[5.6.22] ETSI ETS 300 659-1, Public Switched Telephone Network (PSTN); Subscriber line protocol over the local loop for display (and related services); Part 1: On hook data transmission (February 1997)

[5.6.23] A Közösségfejlesztés magyarországi honlapja

(http://www.kkapcsolat.hu/el_konyv/legalabb/fogyatek.htm)

[5.6.24] ETSI ETR 166, Human Factors (HF); Evaluation of telephones for people with special needs; An evaluation (January 1995)

[5.6.25] ETSI ETR 029, Human Factors (HF); Access to telecommunications for people with special needs, Recommendation for improving and adapting telecommunication terminals and services for people with impairments (October 1991)

[5.6.26] ETSI ETR 345, Human Factors (HF); Characteristics of telephone keypads and keyboards; Requirements of elderly and disabled people (January 1997)

[5.6.27] A Siketek és Nagyothallók Országos Szövetségének a honlapja (http://www.sinosz.hu)

[5.6.28] A MATÁV Rt. honlapja (http://www.matav.hu)

[5.6.29] ITU-T Recommendation P.370; Coupling Hearing Aids to Telephone sets. (08/96)

[5.6.30] ETSI ETS 300 381, Telephony for hearing impaired people; Inductive coupling of telephone earphones to hearing aids, 1994

5.7. Application Programming Interfaces, API-s

Farkas Géza, author

Mazgon Sándor, reviewer

The abbreviation API is very widely used in the world of IT/telecommunication. More then 1000 APIs are implemented in different practical applications, and what is obvious in such a broadly used term, there is more then one interpretation to this abbreviation. ISO, IEC, and the Internet-society defines it as “application program interface” or “application programming interface”. In the ITU usage API appears as “application programming interface” or “application programmatic interface”. Within this suite of interfaces other interfaces can be included, for example the “programming communication interface”. From now on we are discussing the mostly wide-spread “application programming interface”.

The application programming interface forms the interface between the software applications and the telecommunication networks, the CTI (Computer Telephony Integration) is realised through the API. The integration is done at layer 7 of the OSI model. There are a lot of standards in this field, which allow the interoperability among various products made by different manufacturers.

Although ITU-T has developed a lot of relevant standards (e.g. F.581, H.324, I.312, T.171 … T.176, T.180, T.611, Q.1201, Q.1300 … 1303, V.130, X.638, X.446, Y.110, etc.), nowadays mostly other standards are used for interoperability purposes in the practice.

We summarise briefly the evolution of the different standards used in practice:

- ECMA has overseen the CSTA (Computer Supported Telecommunications Applications) standard since 1988, and then the whole family of standards in this area. This standard specifies the operation of switching and computing environments.
- ECTF specified SCSA (Signal Computing System Architecture) in 1993. SCSA specifications apply to voice, fax and data processing boards for PCs.
- Intel and Microsoft developed TAPI specifications for the CTI application programming interface in 1993.
- Novell and Lucent initiated TSAPI development in 1993.

- JTAPI was developed by Sun Microsystems in 1997. Today, JTAPI has become a “de-facto” standard, as well as TAPI or TSAPI.
- ISO/IEC JTC1 working group specified Telecommunication Application for Switches and Computers (**TASC**) at an ITU-T SG11 working group responsible for Intelligent Networking in 1994. The specifications include a TASC general overview, architecture, functional services and management. ITU-T Q.1300…1303 standards are referring to this.

CTI functions

The main CTI functions to which these standards refer and which are realised with the help of these APIs can be classified into three categories: call control, media processing, and customer data management functions.

Call control functions include:

- Call setup and release-related services such as dialing services.
- Routing-related services such as automatic attendant services and alternative routing services.
- Network interfacing services such as tone detection/generation, call setup/release detection, and in-band signalling detection.

Media processing functions includes:

- Voice/fax processing such as voice recording/announcement, voice and fax sending, storing and forwarding.
- DTMF (dual tone multi-frequency) digit processing, text-to-speech synthesis and speech recognition (such as spoken command recognition, speaker verification, etc.).
- Call logging such as on-line recording, monitoring and call accounting.

Customer data management provides personal information management for call parties. This management utilises calling/called identifications to retrieve the calling/called party information from the database and associate the information with the call during its life so the call can be processed in an efficient manner.

In practice the following API standards become the most widespread standards: CSTA, TAPI, TSAPI and JTAPI, that is why we are reviewing them. The other standards are not (yet or already) of such a great importance.

There are two methods of physical and logical connection between a switch and a computer: the so-called “first party call control” and the “third party call control”.

“First party call control” (called as “first party CTI”, too) means an individual solution, when a connection is established at every working place (phone station)

between the switch and the computer. The mapping between the switch and the computer is unambiguous and unique. The CTI software controls only one single phone station.

“Third party call control”, (or “third party CTI” in other words) first of all is meant for working groups. Computers are normally connected to a central CTI-server through the LAN network. The connection between the telephone switching system and the LAN workstations is realised by this CTI-server. The CTI-server controls the switch and not a telephone station.

5.7.1. Standards Review

CSTA

(ECMA TC32-TG11 - Computer Supported Telecommunications Applications)

CSTA is a computer-telephony interface standardised by ECMA (European Computer Manufacturers Association), which defines a “third party” communication between the computer applications and the telecommunication network. CSTA - Computer Supported Telecommunications Applications –standard is specified in a series of documents released by ECMA. The specification concentrates on the demands of the private telephone networks, but it also takes into account the requirements of other private and public networks.

The standard describes the application interface between the switching function and the computing function. The description is abstract, it does not apply directly to the user-network interface or to the network-network interface. Because of it has only an indirect connection with real telecommunication interfaces, it can be used in any environment, the differences among the different existing interfaces are hidden from the CSTA applications. The interface between the user and the network is not subject for CSTA.

ECMA 179 standard (CSTA Phase I)

ECMA 180 standard (CSTA Phase I)

This pair of standards define how the communication at layer 7 of the OSI model is done by the computer supported telecommunication applications (CSTA) between the computer network and the telecommunication network.

ECMA 217 standard (CSTA Phase II)

ECMA 218 standard (CSTA Phase II)

This pair of standards defines the second phase of the computer supported telecommunication applications (CSTA) protocol at layer 7 of the OSI model between the computer network and the telecommunication network.

ECMA 269 standard (CSTA Phase III)

Standard ECMA 269 defines the third phase of the computer supported telecommunication applications (CSTA) services at layer 7 of the OSI model between the computer network and the telecommunication network. This standard is a part of the set of CSTA Phase III standards and technical reports.

All the standards and reports of this set reflect the practical experience of the ECMA members and are based on a pragmatic, common agreement.

The development of this collection began with CSTA Phase I, which contained the CSTA services and protocol standards (ECMA-179 and ECMA-180). In Phase II the collection was completed with ECMA TR/68 Technical Report. This report shows how the CSTA services and events can be used in typical calling situations.

CSTA Phase III complements the Phase II standards (ECMA-217 and ECMA-218) with some major subject and numerous details. This includes the technology based on Versit CTI Encyclopedia (Version 1.0), put to the disposal of ECMA by Versit. Major subjects are the following:

- new services and events categories, such as exchange of skills, call accounting, media processing services, calling data records (CDR), etc.,
- • supplementary call control and device control services and events,
- • enhancement and correction of the existing services and events,
- • organising the services and events in order to reflect the function based grouping (call control, device control, etc.),
- • use of consistent models for services and events; these models include the connection start and end state, the connection state transitions, the event monitoring sequences, etc.

First version of ECMA-269 standard was released in December 1997, the second version in June 1998. This release complements the CSTA services planned in Phase II with the following:

Modelling an ACD and an agent connected to an ACD, call related services, calling data recording services, skill exchange services, data collection services, I/O services, logical device services, physical device services, media processing services, operation events, vendor-specific developments, voice services.

The CSTA Phase IV elaboration has begun in 1999. This phase will support the IP based functions such as VoIP.

TAPI

(Telephony Application Programming Interface)

Microsoft's Telephony Applications Programming Interface (TAPI) was developed for Windows based PCs. First version (TAPI 1.3) was released in November 1993. This version was designed for the „first-party" call control configuration running on a 16-bit processor. An enhanced version TAPI 1.4 was released as part of Window 95. These versions do not support a call model such as the one defined in ITU-T Q.931. TAPI enables a speech/data application to set up and release of calls, monitor progress, detect CLID (calling line identification), perform identification, and activate features such as hold, transfer, conference, park and pickup. It can redirect and forward calls, answer and route incoming calls, and generate and detect DTMF signals. TAPI enables multiple applications to share a single phone line. For example, different types of incoming calls (e.g., voice mail and fax) can be accepted on the same line. TAPI provides access to various telephone network services:

- Plain old telephone service, which supports one type of information (voice or data) per call and one channel per line.
- ISDN, which supports simultaneous voice/data per call and multiple channels per line.
- Digital network services, which support data communications.
- Other services such as Centrex, PBX and KTS (key telephone system).

The later versions of Microsoft TAPI (TAPI 2.0 and TAPI 2.1) have moved from the "first-party" call control configuration to "third-party" call control configuration. The latest version TAPI 3.0 provides a much more friendly and powerful telephony environment. The TAPI 3.0 Component Object Model allows developers to write TAPI-enabled applications in various languages including Java, Visual Basic, and

C/C++. TAPI 3.0 supports IP telephony services that comply to the ITU-T H.323 standard.

With the services provided by Windows TAPI an application developer can complement the applications developed for operation systems supported by Microsoft Win32 (such as Windows NT or Windows 95) with telephone communication. TAPI, the Windows platform, Windows APIs used for other communications and ActiveX drivers forms an ideal base for development and use of telephony applications.

TSAPI

(Telephony Services Application Programming Interface)

Was developed by AT&T and Novell. TSAPI is a Netware-loadable module (NLM) that resides in a Novell server. The software developers implement the applications by using the TSAPI specification so they do not need to directly access the various manufacturers' switch-to-host interfaces. The TSAPI-based switch services are typically created on top of the CSTA switch-to-host interface. Although TSAPI was designed for both "first-party" call control and "third-party" call control, most of its applications focus on third-party call control.

JTAPI

(Java Telephony Application Programming Interface)

JTAPI was developed by Sun Microsystems. First version of it was released in 1997. The actual version JTAPI 1.3 came out in 1999. This standard is spreading rapidly.

JTAPI adds telephony functions to Java programming language. With the help of JTAPI the applications can run on different operation systems and on different hardware platforms.

JTAPI was meant to be just a simple API. It requires of the application developer to know the telephone network, but there is no major need for the implementation-specific knowledge needed for application development.

JTAPI can be used for "first-party" and "third party" call control and media stream control.

JTAPI is not a new telephony API in fact, though it can be implemented without an existing telephony API. It is meant to be built on top of other existing telephony APIs (for ex. TAPI, TSAPI), and so new applications can be developed with it.

5.7.2. Other institutions

Versit

Versit is an initiative launched by Apple Computer, AT&T, IBM and Siemens. These companies developed a series of specifications for computer communication, including CT. The Versit projects include the TSAPI development, the CTI Enclyopedia of specifications (it offers service definitions, protocols, and the TSAPI-based APIs), and the definitions of the vCard Electronic Business Card. The Versit founders joined ECTF in 1995. At the end of 1996 Versit finished its activity, and it is not developing any interoperability specifications, but it is working in other subject.

ECTF

(Enterprise Computer Telephony Forum)

ECTF is an open organisation, established to support and to realise computer-telephony implementations based on “de facto” and “de jure” standards. ECTF defines two types of servers: application servers and CT servers. The application servers execute telephone and media applications in a shared network. CT servers supply the telephone and media resources (lines, speech recognition, fax) needed for the applications. Taking some key-interfaces from the primary system elements and defining them very carefully, it is possible to achieve interoperability in a very wide area. ECTF defined the following interfaces:

- S.100 media and switch services interface
- S.200 transport protocol interface
- S.300 service interface
- H.100 hardware compatibility interface
- M.100 administrative services interface

Call control and other interfaces providing application interoperability are under development.

For example ECTF has been dealing with JTAPI development since 1997, too.

6. Editorial remarks

Chapter 6 of our (system-)reference book – in the spirit of the hierarchical structure of telecommunication systems – gives an overview of applications that are, or could, be based on the technical possibilities of lower stages (described in the previous chapters).

Due to the nature of the topic the reader will find – instead of a close, disciplinarily interdependent approach– a medley of applications which (quasi) illustrate or "expand" the "space" of applications.

The fast development of telecommunication and its convergence on informatics (IT), as well as digital media, has opened a wide, quasi unlimited number of application possibilities, ranging from technical ones to usage in arts, from public aimed discussions to army adaptations. In some cases telecommunication accelerates and makes human communication easy and independent of place. In other cases it broadens traditional human activities with new possibilities (e.g. tele-medicine, distance learning), which react basically to (backwash) the paradigms and technologies of the given scientific field, opening up new areas several times. For example, distance learning does not simply facilitate access to materials to be learned, but also creates new didactical, experimental theories, moreover it transforms the whole human knowledge transfer process in the long run. It changes the participants in the transfer of knowledge, defines new teacher roles; the number of participants and the structure of the formation will also change, the aims of the students will be different. The "school" itself will be different. In that case, for example, it is difficult to decide what is at all important in an introductory description. Is it the list of requirements and arguments for organising telecommunication networks (type of network, loading, terminal type, protocol etc.) or the effects of change in paradigms caused by telecommunication systems and their predictable trends, or the presentation of the effects of a new teaching system on the personality, or perhaps the economic aspects of the system? Finally, human activities are often organised through telecommunication networks that simply did not exist in the past (e.g. tele-house communities, e-mail communities, new ways of practising personal rights directly, etc.).

Since then, because of the possibility of making multiple connections, the acceleration of the innovative activity of mankind is exclusively due to the telecommunication networks. Applications realised in practice are only limited by human imagination, thus the appearance of newer and newer applications is a dynamic, accelerating process. However, the consequence of this is that there is often no real

experience or empirical background in the elaboration of challenging possibilities of huge public interest, or there is no accepted terminology for their description. This is especially true in the case of a small country with a population of 10 million where limitations of the number of affordable and analysable projects are evidently palpable. This difficulty of the presentation of applications is increased by the fact that the application fields in general are independent disciplines of such an extent that their presentation would require a book of similar length to the present one.

Because of the above difficulties, under the pressure of necessity, we have consciously chosen illustrations in this chapter of our book by arbitrary selection. Further, we have had to accept with a heavy heart how many already mature, well tested, important areas are not included in our "book". For example, there is no mention of remote data collection or the issue of remote control, so the control of intelligent buildings remains unmentioned, and also because of this we have not mentioned "telejamming" (distant co-working, co-operation, e.g. distance music playing together), or co-operative "virtual games" that put more and more emphasis on strategic action control. But distance "meta"-questions have not been treated either, which form a substantial part of the problem: how will increasing communication possibilities accomplish the task of broadening human dignity. So we have not spoken about the truthfulness of information available on nets, the ownership and price of published contents, about the gathering and exploiting possibilities and limits of data relating to individuals.

The evaluation, language, depth and aspects of presentation, concerning the questions chosen for presentation, differ from each other, their integration would exceed our possibilities. Moreover, because depending on where the telecommunication nets are used, their aim, social atmosphere, the intelligence of the users, and the historical roots of the basic discipline differ– to be brief, the language of discourse is different in each field.

What was presented in the case of each example, apart from the things mentioned above, depended largely on the personality and experience of the author of a subsection or the career he/she has followed. What has been defined, illustrated from an area also depends on these. In connection with this, sometimes the capacity requirements of telecommunications channels were primarily presented, and sometimes the sociological typology of user applicants was given emphasis.

So, based on the thoughts outlined above we have tried to give examples, which show from many aspects, cross areas of the present possibilities. We have shown applications which offer firstly the expansion of human communication possibilities, as well

as demonstrating services that facilitate the usage of communication tools or make it more useful. One can find such examples in subsections 6.2 – 6.4.

In another groups of applications we have illustrated existing areas of human activities which are basically influenced by the new communication technology. Besides outlining the problems and new possibilities within of these areas we have also aimed at the questions of implementation connected with the net. One can see examples in subsections 6.5 – 6.7.

The third group of the examples shown illustrates new special areas i.e. activities whose existence is simply the consequence of the new communication technology. In these cases we have put the emphasis on the illustration of the given phenomenon, its development and social background. Such subsections are 6.8 – 6.9.

Finally, when here, in the introduction to this chapter, we ask the reader's understanding at the same time we venture the prediction that probably this chapter of the book will be the least permanent part of it. Prospectively, this subsection could exploit, at most, the on-line character of our book. This also means that one can find the largest number of open questions on these pages, questions whose solutions allow for human invention, including also the reader's initiative. After all, we know from practice that fundamental research and development of basic technologies are mostly going on in a few big, capitalist countries. The determinant role of smaller participants is the application of results their usage with imagination Here can be found the area that is usually mentioned as a gap-technique reflecting the future of Hungary, i.e. finding the research and development activities not cultivated by the big ones, which, on the one hand, rely on intellectual forces of the country, but on the other hand, whose solution promises economic profit as well.

Miklós Havass, Editor of the Chapter

6.1. Overview of applications

Dr. Péter Bakonyi, András Ercsényi authors

György Takács PhD, reviewer

Over the last decade the dynamic development of technology has created new possibilities which have an influence on economic as well as social development. Information and communication technologies are an integral part of everyday life and provide useful services which can be used at home, at work and in various other situations.

There is a new society on its way: the emergence of the information society is in close connection with the demand for the opportunities provided by the new technology. The information society is not a utopian vision of the future, but one that we already live and exist in. New generation applications emerge from the intergration of telecommunications, informatics and the media, which change our work process, our way of life as well as social communication.

These so-called „new generation" applications are the topic of Chapter 6. These applications cover nearly all areas of economy and everyday life. At the same time, it must be noted that these new generation methods have brought up and keep bringing up several problems. Most important of these problems are the ethical, cultural and legal issues. Due to the globally widespread nature of the media the controlling power of national governments has nearly disappeared thus administration and regulation rights have become part of the private sphere. (Internet service providers, who are in charge of running the system of the World Wide Web, are in private ownership in nearly all parts of the world!) The appearance and spreading of the Internet has shown the need to reconsider several issues, such as the free flow of ideas and the free expression of opinion. The World Wide Web has become the symbol of liberty – while it seems that the necessary control over it can hardly, if at all, can be implemented.

Legal regulation is a similarly significant issue, involving the protection of the private sphere, the question of responsibility and even Internet-crime. The latter does not only mean the illegal entering of computer systems via the Net, but also the use of the World Wide Web by terrorists. A summarizing overview of legal, cultural and administration

issues in connection with new generation applications can be found in Chapter 6.7: *Administration, culture and law*.

Following is an introducion of the typical applications and an attempt to systematize these.

The new generation of business activities and business services is electronic commerce (e-commerce) or also called tele-commerce. This is a type of business activity where the partial activities are based on eletronic processing and transfer, and where the World Wide Web (shortly web) technology plays a significant role. Electronic commerce , among many other activities, includes the eletronic distribution of products and services, the electronic use of electronic money and the electronic customer services.

Electronic commerce can be considered the new form of business activities which provides a new and effective means of getting the products into the market, and where the electronic handling of customer service activities also plays an important role.

In the more simple versions of electronic commerce, transactions such as the purchase and sale of products and services are only partially carried out by an electronic medium. The aim, however, is to implement the exchange of goods, products, services and contents, as well as the realization of financial transactions with the use of electronic devices and technology.

In the expression electronic commerce, the adjective „electronic” refers to the fact that the participants of the business transaction connect through an electronic medium, that is, a network of computers. Network connection usually means distant access – although people working next to each other in the same office are connected in the same way as distant partners. This type of distant connection can also be indicated with the Greek prefix „tele”. therefore the expression „tele-commerce” can also be used instead of electronic commerce. Words such as tele-medicine or tele-cure are of the same root, indicating the new generation medical application where the doctor can access an X-ray or an electrocardiogram through the Net.

In Hungarian the prefix „tele” is also used for new applications. This way tele-education or tele-working stand for a new electronic type of education or employment where the use of the Web becomes necessary through a computer network.

Chapter 6.5 *Online services, e-commerce* describes the topics in greater detail.

With the spreading of computer network technology, distance education and learning (e-learning) enabled the direct distribution of educational materials based on

individual needs to the appropriate places, the appropriate people and at the appropriate time. The target community is one including consumers in higher education, business as well as individual consumers. The educational method is interactive and user-centered. E-learning is divided and uses the XML technology. XML is a data form developed for structured or partially structured texts directly shown by different media. Today XML is a de-facto data exchange standard used by not only distance education systems, but nearly all „e" systems.

New generation medical applications have also appeared, such as new generation computer systems in hospitals, developed tele-diagnosis and consultational systems. These systems use medical networks which enable the improvement of medical services and the higher participation of patients in preserving their health. There are seperate systems for supporting the disabled as well.

One of the significant challenges for the information society is to provide access to as many people as possible to the cultural and intellectual heritage found in public museums and libraries. Digital multimedia technology enables the forwarding of information to an extensive layer of consumers as well as interactive access to this information. Thus, with an appropriate infrasturcture the above mentioned cultural values can become available for the public.

Information society offers new solutions to various transport-related problems. Through the various telematic services the safety of the road system, traffic on the roads and orientation as such shows significant improvements with the use of navigation systems which are built into the vehicles. At the same time the capacity of the roads can also be optimized with the new navigation systems. Pollution might also decrease with the new services.

All of these issues are dealt with in greater detail in Chapter 6.6 *Tele-services, tele-medicine and tele-working*.

Tele-houses or also called community tele-service centers are becoming widespread mainly in small towns. As we can tell from the name, these are public sevice units which are connected to the infocommunication system in a such way that the use of the devices available at the tele-house is assisted by a professional staff.

Tele-houses are service provider institutions equipped with information and communication technology. Their duty is making infrastructure available for all purposes which are vital to the small community. Such purposes can be providing access to public

administration, the use of libraries, distance learning, electronic correspondence or possibly tele-working.

Tele-huts are smaller, developing tele-houses with less devices and services.

The implementation of tele-houses as multifunctional tele-service centers varies significantly all over the world. The tele-house movement does not only exist in underdeveloped or developing countrries but in developed ones as well. The local version is the so-called Hungarian model. According to the general opininion, tele-houses might indicate the endpoint system or the beginning of the information public network.

The effective operation of tele-houses requires the collaboration of the local community. If successful, the town, the people (employees, enterpreneurs, customers and salespeople) can become part of the electronic economy. Chapter 6.9 *Tele-houses* deals with this intriguing topic in greater detail.

Telecommunication, or also called telematic terminals are devices which provide e-services to the general consumer. Such devices appear in every place where people as consumers connect to some sort of service provider. A good example for the telematic terminal is the so-called integrated home multimedia terminal or the large community information terminal. An example for the former is a television which provides intergrated Internet functions. (Nowadays this is considered a rarity, since only the reverse is used at most, which is the TV card built into the personal computer.) Large community information terminals provide information for non-professional consumers at exhibitions, local authorities, tourist centers, etc.

One feature of telematic terminals is that their use does not require extensive professional knowledge. Their make-up is application-oriented. Chapter 6.4 *Telematic terminal* discusses this topic in greater detail.

Among new generation services speech or voice transfer through the (communication) network is of particular importance. This technology is called VoIP, stemming from the English expression „Voice over IP", that is, voice transfer over the IP[1] protocol.

VoIP emerged as a result of the so-called communicaton convergence. The main purpose of this convergence is to ensure that speech transfer, traditionally belonging to the

[1] IP or Internet Protocol is a dialogue regulation called „Internet". It is a confusing coincidence that one of the dialogue regulations of the Internet World Wide Web is also called „Internet". This protocol is responsible for ensuring the errorless transfer of data between the end-points of the same type of networks (such as the Ethernet or the token-ring network).

field of telephony and data transfer, traditionally belonging to computers are provided as the service of one sole system. Nowadays convergence, apart from speech transfer has taken on other meanings, according to which wired and wireless transfer systems are also heading towards a uniform solution.

The technological basis of convergence is IP, considered as the common transfer solution. The success of IP stems from the fact that it has not defined the sub-system it is based on, and it is simple enough to function on any medium acting as sub-system.

The process of convergence is in full swing: the industry of telecommunication is undergoing a significant change, of which we sometimes become a suffering part, when we have to replace our mobile phone with another one.

Further areas of the convergence are looking for ways to implement uniform transfer of messages, where faxes, e-mail messages and voice messages can be handled through the same interface.

IP as a basic transfer technology entails other changes, since it establishes the so-called virtual private networks. These are typically computer networks which are operated by a company with several premises – above the IP system of the public World Wide Web.

Chapter 6.2 *New generation services (VoIP)* deals with this interesting topic in further detail.

High-speed computer networks and high quality voice transfers can only function with broad band signal transfer systems. Modern cable television (CTV) networks provide a technically high capacity and at the same time economical infrastructure for broad band applications. The question arisies: how does CTV network relate to Internet systems? It is also interesting to consider whether there is convergence in this field, that is, when do we get from CTV networks to interactive digital networks? The process has already started and the approach of the two networks is indicated by high quality Internet access through cable television networks.

The fact that, based on the Internet-telephone model, there are various cable telephone services available shows the expansion of CTV systems. All of these are the forerunners of the introduction of various new tele-services.

The direction of further improvements of CTV networks has already been outlined, the essence of which is that so-called „full service” networks will replace today’s program distributing networks, that is, the era of interactive digital television watching will begin.

Chapter 6.3 *Cable television as multi-service media platform* discusses thid topic in further detail.

Tele-working in the USA and several Western-European countries already has its traditions. The necessary requirement for the introduction of tele-working is the technical background based on the spreading of telecommunication devices and the appropriate knowledge of their use. Furthermore, the technical requirement of tele-working is the already existing or newly set up network connection between employer and employee.

In some industrial branches tele-working is a highly beneficial solution. In the USA, for example, there was a shortage of professional manpower due to the dynamic improvement of the software industry, which could not be settled in the traditional way. Tele-working, however, provided a solution to the situation.

In the area of the European Union 6% of the employees are involved in tele-working on average. According to statistics, these employees are between the ages of 30-40, have a university degree and are (for now mostly) men. They do not live more than 50 kilometers away from the employing company, therefore constant personal contact is also provided according to individual needs. There are great number of other characteristics of tele-working, for example, it significantly changes the system of professional personal contacts, a topic on which various sociological essays have been written.

In Hungary one of the best propagandist for tele-working is the tele-house movement, which is more than necessary, since the widespreadness of tele-working has not yet reached the desirable level.

This topic is elaborated in Chapter 6.8 *Tele-working*.

6.2. Next Generation Services

Gábor Rét, author

Tamás Montvai, reviewer

This section gives a brief overview on recent developments in convergence of voice and data communication. The first part of chapter 6.7 deals with effects of convergence and the global emergence of the Internet on telco carriers and service providers. We also give a number of examples, what kind of enhanced services should telco companies (both next gen and incumbent) roll out in this rapidly changing world, in order to stay competitive. Then, in the rest of the chapter we go into a bit more detail on five of those enhanced services: Voice over IP, Virtual Private Networks, Unified Messaging, Presence, and Content Delivery.

6.2.1. Communications Convergence

The telecommunications industry is in the middle of a fundamental shift – one that will dramatically change the way the world communicates. At the center of this shift is the movement to converge voice data and wireless services onto a single network. As a result, service providers and carriers are facing new challenges and increased competition as well as tremendous opportunity.

At present, many communications service providers are forced to maintain two and sometimes three separate networks: the voice (circuit-switched) network, the data (packet-switched) network and the mobile phone (wireless) network - each of them costly to build, develop and operate. As carriers and service providers face falling long distance calling rates and declining margins due to technology innovation and intense competition, they must become more flexible in operating, provisioning, maintaining and administering their networks and must develop innovative offerings that will attract and retain customers.

A. IP is the Common Underlying Technology

The next generation infrastructure is a converged network where one underlying technology is used to carry all types of traffic - voice, data or video. Internet Protocol (IP), a packet-based routing technology, has been chosen by the communications industry as the

universal transport for converged networking due to its lower cost of equipment and ease of provisioning, among other reasons.

Worldwide revenue from voice services has far outpaced that from data services - a trend that is expected to continue. This highlights the importance of voice over IP (VoIP) services. Companies from various segments of the communications industry, including Cisco and AT&T have publicly announced their strategic focus on VoIP.

The first “killer” application to emerge for VoIP was long distance arbitrage. Addressing a market demand for cheaper long distance calling, start-up ventures began to offer free long distance phone calls from PCs and Internet phones. Subsequently, a number of companies referred to as Internet Telephony Service Providers (ITSPs) or “Next Gen” Telcos capitalized on this trend and provided the technology and services that allowed users to make low cost phone-to-phone long distance calls using VoIP. This led to the explosive growth and competition of companies using VoIP to undercut traditional long distance rates.

Today, long distance arbitrage is still the most widely used application of VoIP. However, cheap dialtone has quickly become a commodity. It is increasingly difficult to maintain attractive margins as customers shop and compare for the lowest rates, generating little loyalty to any given service provider. To achieve higher profits, grow market share, and retain customers, telecommunications providers must differentiate services and offer more than dialtone.

B. Next Generation Applications and Enhanced Services

A VoIP enhanced service leverages the benefits of IP to extend communications with new offerings that bring incremental value to end-users. These services deliver new capabilities, greater control over communications, increased efficiency in placing and receiving calls, and the ability to merge the functionality of multiple communications devices.

A VoIP enhanced service can extend the long distance offering of a service provider and clearly differentiate them from the competitors. For example, offering a PC-based application that can more effectively manage and more easily reach business and personal contacts would add value for subscribers. With a more distinctive service offering, existing customers will no longer choose vendors solely on long distance prices, but will also consider the value that enhanced services bring to their communications. Such competitive differentiation promotes customer retention, as providing VoIP enhanced

services increases customer satisfaction. Finally, enhanced services can drive incremental minutes on a service provider's VoIP network, delivering higher margin network services across a fixed investment.

Today's evolved VoIP infrastructures make it possible to deliver true enhanced service capabilities. VoIP standards such as H.323 and SIP have become widely accepted. Carrier-class VoIP gateways, gatekeepers, proxy servers, and softswitches have already been deployed in networks and have addressed telecommunications providers' requirements for scalability, availability, manageability, and call control.

The factors that account for the popularity of the World Wide Web are the same factors driving the demand for enhanced services. Users show interest in solutions that simplify their lives, reduce the volume of their tasks and provide a diverse range of meaningful functionality. In deploying next generation enhanced services, success will stem from providing integrated solutions that benefit customers with added ease of use and increased efficiency and productivity.

When defining and deploying enhanced services, it is also important to focus on extending existing user behavior rather than expecting to change it. Users associate telephones with placing and receiving calls, while the PC is associated with messaging and contact information. Therefore, VoIP enhanced services should *give users the option* to place and receive calls from their phone handset or mobile phone and view messages and contact information from a PC monitor. By embracing these preferences, not fighting them, service providers can accelerate user adoption.

When defining and developing enhanced services, providers should take advantage of the intrinsic benefits of a converged voice and data network versus a circuit-switched network. In fact, the trend is not only that *networks* converge, but that *applications* converge as well. A VoIP enhanced service can be differentiated by:

- Providing a choice or combination of PC and phone endpoints.
- Leveraging the more powerful PC user interface for management of contacts, availability and messaging.
- Having a PC interface that is customizable and appropriately branded.
- Using a web-based PC interface to schedule (and potentially, to moderate) phone conferences.
- Embedding click-to-call capabilities within web pages.
- Initiating phone-phone calls from a PC application; from which higher value long-distance and conference calls could be made from a "captive" audience.

Examples of enhanced VoIP services that are deployed today or planned for deployment include:

VoIP Conferencing

This enhanced service offers IP-based voice conferencing that can be scheduled and managed from a PC. Often, conferences can be initiated from a phone as well. This type of service application provides the means to conduct and moderate calls from a PC or a phone conference. Conferences can be set up to be scheduled “meet me” style meetings or ad hoc conference calls. The benefit of VoIP conferencing is the ability to include substantially more voice conference participants than can be supported by an office PBX system, at a price much lower than what can be offered by traditional conference bridge service providers. Because of this, VoIP conferencing can provide compelling add-on capabilities to an IP-PBX or telephony-enabled LAN deployment.

VoIP Calling

Sometimes referred to as “eCalling”, VoIP calling provides for inbound and outbound calling over IP. Calls are initiated from PCs, but participants (and even the call initiator) can use standard PSTN telephones and mobile phones as the VoIP communication device. VoIP calling applications combine presence detection with click-to-call telephony capabilities to improve call completion ratios and user productivity. Callers can determine availability of intended call recipients and complete calls to a PC or phone. The core benefit of a VoIP calling service is greater productivity. Subscribers can reach the contacts they want faster. The benefit to service providers is more proactive calling and conference calls, driving usage on their VoIP networks.

Web Based Click-to-Call

A click-to-call application can be provided as a service to businesses with e-commerce web sites and call centers as an additional channel for customer communication. Online customers click a button on a Web site to initiate an IP call to the company’s call center agent. In addition to PC-to-phone and PC-to-PC connections between customers and agents, PSTN callback is typically made available for customers who are not equipped with PC microphones. Web-based click-to-call provides immediate interaction with customers, leading to higher sales revenue. Agents can respond to customer inquiries immediately and leverage upsell opportunities. In addition, agents will

be more productive and efficient, as less time is required to address customer needs and close sales.

Record and Playback of Voice Session

This enhanced service allows subscribers to record voice conversations, whether they are point-point calls or conference calls. Ability to record a call is given to the call initiator or conference host and is available via both phone and PC. Playback of the voice session is accessible by any phone and PC user, if given access to the recording by the initiator. Record and playback, especially as an option for conferencing services, provides participants the ability to review previous discussions. For online education, it gives students who have missed a session an ability to listen at a more convenient time. For call centers, record and playback capabilities can ensure agent service quality or accuracy of customer transactions.

Find Me / Follow Me

This service allows subscribers to provide one phone number to all of their callers. A subscriber can screen or receive calls (on a PC or phone) depending on who is calling, time of day, or other criteria. Subscribers are given the flexibility to have any of their communication devices ring based on their preferences, regardless of the number that was dialed. Find me/follow me provides the benefit of making subscribers more reachable.

Instant Voice Messaging

This enhanced service allows subscribers to broadcast short instantaneous voice messages to a group or an individual. Typically, messages can be sent and retrieved from either PC or phone. This type of capability is a convenient addition to voice mail. It is particularly powerful when combined with services that provide group directory services.

VoIP Notification Services

A notification service automatically calls a subscriber’s phone and delivers a short voice message, allowing the subscriber to interactively respond via DTMF tones or by initiating a real-time VoIP call with a service agent. This application is particularly useful for delivering time-sensitive information that a subscriber would act upon (for example, stock price quotes or airline flight rescheduling). Businesses who adopt this service can engage their customers more proactively, and as a result increase sales and customer satisfaction.

VoIP notification services are another example of how service providers can create more value in a phone call for their customers.

Presence Detection

Presence detection capabilities offered in conjunction with VoIP conferencing or click-to-call functionality provides users the ability to confirm availability of a contact and easily complete a call using VoIP or PSTN end points. Knowing someone's availability before attempting to contact them greatly increases the chances of completing a call. By embedding availability in e-mail signatures, corporate directories, contact management tools or anything else that supports the projection of an HTML object such as a WAP phone, callers can determine availability and complete calls to PCs or Phones.

Unified Messaging/Unified Communications

Unified messaging (UM) combines voice mail, fax and e-mail into a single application for storing and retrieving an entire suite of message types (for example, WAV files for voice mail and TIFF files for fax). Through a common interface, users can access these unified services using PC, phone, or mobile device. Unified communications takes UM services and extends them. Unified messaging is an asynchronous, "store-and-forward" type of service. A unified communication service integrates UM with real-time VoIP communications and presence capabilities. For example, an e-mail, page, or fax message retrieved via PC or phone can be replied to with a real-time VoIP call to the sending party. Other capabilities such as Internet call waiting and number/mailbox consolidation are often included in unified communications offerings. Unified communications as an enhanced service delivers cost-savings, convenience and increased productivity. Service provider infrastructure costs are reduced, as there is only a single application to support. Productivity is increased, as users no longer need to check multiple systems for e-mail, voice mail and faxes. Presence information allows users to know if message senders are available for a conversation. Click-to-call capabilities allow PC users to immediately reach available contacts for a VoIP call.

6.2.2. Voice Over IP

Voice over IP (abbreviated as VoIP) is the transport of voice over the Internet Protocol. In simple terms, it is the ability to make a telephone call over the IP network, at a cost much lower than traditional telephone networks. Even though VoIP presents a

tremendous opportunity, the Internet Protocol was not designed to carry voice, and hence providing toll quality voice over IP is a major challenge. Major issues revolve around the quality of the voice calls as well as the ease of use for the end user. Significant progress, however, has been made in this respect to engineer the packet networks to provide toll quality voice. VoIP hence has a major potential for being a low cost alternative to PSTN. It also has the potential of replacing the telephone network with an integrated network, capable of supporting both voice and data over a common infrastructure. In the next few pages various aspects of VoIP are discussed, including the challenges, and the related standards.

A. Issues and Challenges in VoIP

QoS

The biggest problem faced by voice over packet networks is that of providing the end users the quality of service that they get in a traditional telephony network. Unlike the PSTN, where a dedicated end-to-end connection is established for a call, packet based networks use statistical multiplexing of the network resources. Though sharing resources amongst multiple users leads to a cost saving, it deteriorates the overall quality of service offered to a user. There are multiple parameters that determine the quality of service provided by a network. These include delay, the delay jitter and the packet loss experienced in a network.

Delay - The delay experienced in a packet network is classified into the following types: the Accumulation Delay, the Packetization Delay and the Network Delay. Each of these adds to the overall delay experienced by a user. The accumulation delay is caused by a need to collect a frame of voice samples for processing by the voice coder. This delay depends upon the sample time and the type of voice coder used. The accumulated voice samples are next encoded into a packet, which leads to the Packetization delay. Finally, once this packet is sent through the network, it experiences some delay to reach the destination. This is caused because of multiple factors, which includes the processing done by each intermediate node in the network to forward the voice packet, the capacity of the underlying physical medium, etc. Delay in transporting a voice packet through the network leads to two main problems, namely Echo and Talker Overlap. Echo becomes a major problem when the round trip delay through the network becomes greater than 50 milliseconds. To avoid the problems of echo, echo cancellers are required. Since packet networks introduce higher end-to-end delay, and hence have a greater round trip time,

echo cancellation is an essential requirement for a voice over packet network. Apart from this, in case the end-to-end delay becomes greater than 250msec, the problem of Talker Overlap surfaces. This is experienced by one talker overlapping the other talker, and can be extremely annoying. Also, a delay of more than 250msec feels like a half-duplex connection and cannot be claimed to be an interactive session.

Jitter - Two packets between the same source and destination might experience different processing delays and different congestion situations in the network, resulting in a variation in the overall delay experienced by the packets. This variation in the delay experienced by packets is measured as delay jitter. Also, this might lead to packets reaching the destination out of order. Unlike data packets, voice packets would be severely affected by the delay jitter. To take care of the jitter, a buffering scheme is used at the destination. Packets at the destination are received and buffered. After the buffer is full to a threshold value, the packets are played in sequence and with a constant delay, minimizing the delay jitter. However, this buffering of packets at the destination leads to an additional delay and also adds up to the other three types of delays discussed above.

Packet loss - Since IP is a best effort protocol, packets routed through an IP network can be lost. To provide reliable transmission of data in an IP network, a retransmission scheme is used at the transport layer, which retransmits any packets for which an acknowledgement is not received from the destination (assuming that the packet got lost). However, the same scheme cannot be applied to voice, as by the time a retransmitted voice packet reaches the destination, it might be useless. To compensate for lost packets, a few techniques are specified for voice over IP networks. Of these, one scheme proposes to play the last packet received again, to compensate for the lost packet. However, this scheme cannot work if there is a burst of lost packets. Another scheme proposes to send redundant information, increasing the bandwidth requirements. Thus, lost packet compensation is an essential requirement for supporting VoIP.

Billing and Interworking

Multiple variants of Voice over Packet Networks exist today. Amongst these is Voice over ATM, Voice over IP and Voice over Frame Relay. All these variants would have to co-exist with each other, along with the conventional PSTN. This raises multiple issues, but primarily the issue of Interworking. Let us consider a scenario where a user in an IP network places a call to a user in a normal telephone network (PSTN). Then the call would have to traverse through a VoIP network to the PSTN through a gateway, which would

have to provide the necessary interworking, including address translation etc. Also, billing becomes complex in this scenario, since the call traverses networks that differ in principle and also in charging scheme. Needless to say, error localization and reporting would also become more complex in such a heterogeneous scenario.

Transparency of Operation

As discussed above, the underlying mechanism supporting a voice call could be different, based upon the type of network (VoIP, VoATM, VoFR, PSTN, etc.) What is important in this scenario is to provide the user with a feel of a conventional telephone network. The call establishment process (call control and the associated signaling process) must be made transparent to the user. Also, a user placing a call from a PSTN telephone to a VoIP network, for example, should be unaware of the fact that his call is traversing two dissimilar networks.

B. VoIP related Standards

The standardization activity of VoIP is being governed by two bodies, the ITU-T and the IETF.

ITU-T Standards

The first set of standards related to VoIP was developed by ITU-T through their H.323 series. This standard, along with other standards developed by ITU-T is detailed below.

H.323 standard provides an infrastructure for audio video and data communications over packet based networks that may not provide Quality of Service (QoS). The H.323 standard is a part of H.32x protocol family, that includes, besides H.323, standards like H.324 (standard for multimedia transport over SCNs) and H.320 (standard for ISDNs) among others.

The H.323 standard describes four key components for an H.323 system, namely the terminals, gateways, gatekeepers, and Multipoint Control Units (see Figure 6.2.1). These components are described in the following subsections.

A terminal is a PC or a standalone device running an H.323 protocol and the multimedia applications. A terminal supports audio communications and can optionally support video or data communications. The primary goal of H.323 is to interwork with other

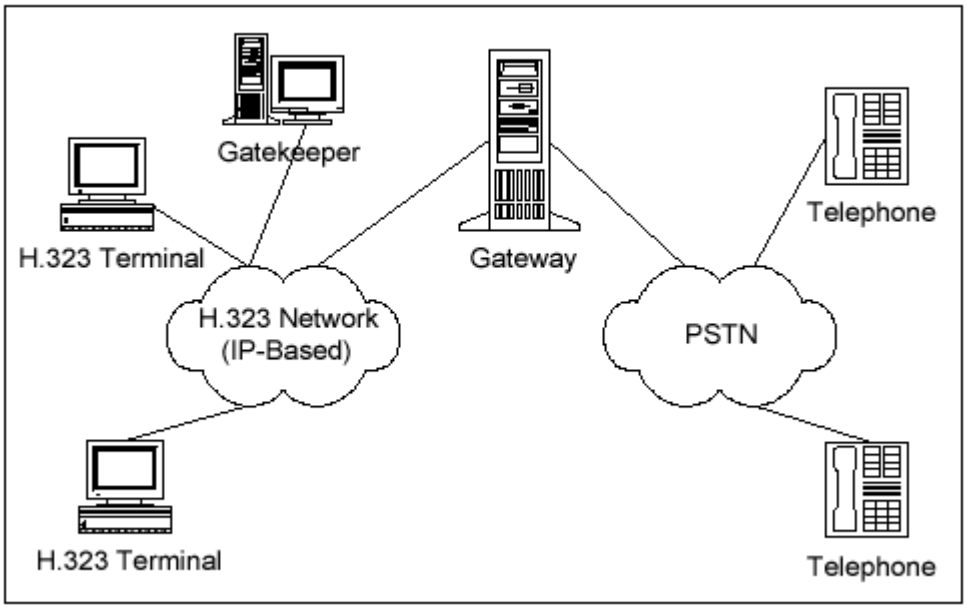


Figure 6.2.1. H.323 Network Components

multimedia terminals. Because the basic service provided by an H.323 terminal is audio communications, a H.323 terminal plays a key role in IP-telephony services.

A gateway connects two dissimilar networks. An H.323 gateway provides connectivity between an H.323 network and a non-H.323 network. This connectivity of dissimilar networks is achieved by translating protocols for call setup and release, converting media formats between different networks, and transferring information between different networks connected by the gateway. A gateway is not required however for communication between two terminals on an H.323 network. The gateways perform functions like search (conversion of called party phone to IP address), connection, digitization, demodulation, compression/decompression, and demodulation.

Gatekeepers can be considered the brain of H.323 network. It is the focal point for all calls within the H.323 network. Although they are not mandatory, they perform important services like address translation, admission control, bandwidth management, zone-management, and call routing services.

Multipoint control units (MCU) These provide support for conferences between three or more H.323 terminals. All terminals participating in the conference establish a connection with the MCU. The MCU manages conference resources, negotiates between terminals for the purpose of determining the audio or video coder/decoder to use, and may handle the media stream.

H.225 is a standard, which covers narrowband visual telephone services defined in H.200/AV.120 series recommendations. It specifically deals with those situations where the transmission path includes one or more packet-based network, each of which is configured and managed to provide a non-guaranteed QoS. H.225 describes how audio, video, data and control information on a packet-based network can be managed to provide conversational services in H.323 equipment.

H.248 is the same as MEGACO published by the IETF, and is discussed bellow.

IETF Standards

Internet Engineering Task Force (or IETF) along with ITU-T is playing key role in VoIP-related standardization effort. The following subsections describe briefly key areas and their standards (RFCs).

Media Gateway Control Protocol (MGCP)

The Media Gateway Control Protocol or MGCP implements the interface between a Media Gateway (MG) and a Media Gateway Controller. This interface is implemented as a set of transactions. The transactions are composed of a command and a mandatory response. The MGCP protocol is detailed in RFC2705.

*Megaco/H.248 i*s the media gateway control protocol defined by the IETF and the ITU-T and is used in a distributed switching environment. It is designed as an internal protocol within a distributed system, which appears to the external world as a single VoIP gateway. Internally, the architecture is designed such that the intelligence of call control is outside of the gateways and handled by external agents (see Figure 6.2.2).

Megaco thus divides the media logic and the signalling logic of a gateway across different functional components. While the Media Gateway (MG) handles the media logic part, the Media Gateway Controllers (MGCs, or Call Agents) control the Media Gateways to establish media paths through the distributed network. An MGC can control multiple MGs. On the other hand, one MG can register with multiple MGCs. Communication between these two functional units (MG and MGC) is governed by the Media Gateway Control Protocol (or Megaco). Megaco is thus a master/slave protocol, where the call agents act as command initiators (or masters), and the MGs act as command responders (or slaves).

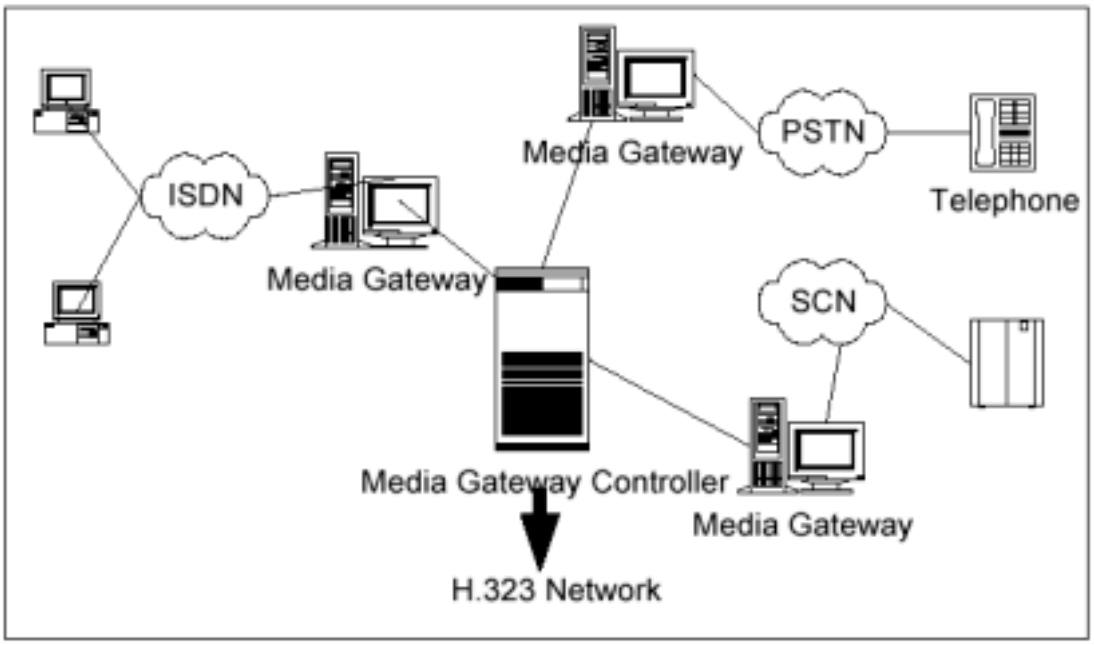


Figure 6.2.2. MEGACO Network Architecture

SIP

Session Initiation Protocol, or SIP, is a communication protocol used for a number of applications like Internet telephony, Call- forwarding, Multimedia conferencing, Terminal- type negotiation, Caller/callee authentication, and a host of other multimedia services. The standards for SIP have been developed by the MMUSIC group of IETF. SIP is transported typically over the connectionless UDP protocol. UDP is preferred over TCP because of its lower state-management overheads, real-time characteristics, and better performance. The standards for SIP include RFC2543 (Session Initiation Protocol), RFC2327 (Session Description Protocol) and a number of Internet drafts, that are being worked upon. Figure 6.2.3 shows network components and sample message flows for SIP-based network.

RTP and RTCP

RTP or Real Time Protocol provides support for applications with real-time properties, including timing reconstruction, loss detection, security and content identification. RTCP provides support for real-time conferencing for large groups within an Internet, including source-identification and support for gateways (like audio/video bridges), and multicast-to-unicast translators. RTP can be used without RTCP. RTP is standardized in RFC1889. Besides this RFC, there are several other RFCs, which discuss specific problem areas of RTP.

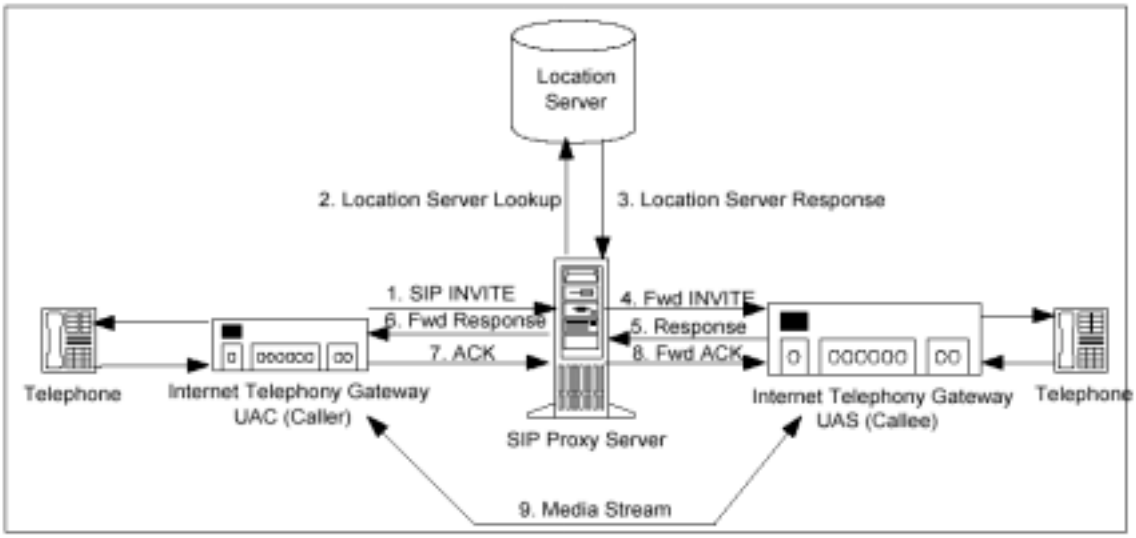


Figure 6.2.3: SIP Network Components and Message Flows

6.2.3. IP-VPNs

From a research report published by InfoTech:

- In the U.S., the market for Managed IP VPNs is projected to grow from $1.5 billion in 2001 to over $11 billion by 2005.
- Convergence VPNs represent less that 5% of the Managed IP VPN market, but by 2005, they will mushroom to 75%.

The term IP-VPN includes two basic models, the CPE- and the network-based IP connectivity. Each has strengths relative to the other: CPE-based solutions make low-cost, secure remote access possible. IP-class-of-service, performance and multi-service integration are better supported by network-based solutions. This latter model of IP-VPN handles data at levels previously only available with Asynchronous Transfer Mode (ATM) and Frame Relay.

By pure definition, a virtual private network (VPN) is the interconnection of multiple sites via a private backbone network, but this term has become muddled, and VPN has been applied to virtual networks based on the public Internet. (The term "IP-VPN" does not necessarily connote data transmitting the public Internet, though some assume as much.)

The definition of IP-VPN has always been soft. For many service providers, it has been the "P" in VPN that was overstated - they put firewalls at the customer sites and send traffic between them via the public Internet. This obviously is less then "Private". IPsec added additional privacy and security, and has proven to be a popular means of connecting two sites via an encrypted tunnel across the public Internet. Recently the IP-

VPN definition has evolved beyond "IP-based networking over a service provider's wholly owned IP backbone network". Today this is known as a network-based IP-VPN. IP-VPN also includes using the Internet to carry traffic, while employing firewalls, digital certificates, encryption, and user authentication to make the communication safe. This firewall-and-Internet-based IP networking combination is known as a CPE-based IP-VPN.

Both forms of IP-VPN have evolved, and each has certain advantages over the other. There are business and technical drivers pushing the growth of IP-VPN technology:

1. the popularity of the IP and the wide acceptance of the Internet,
2. the potential cost savings associated with using IP rather than multiple networking protocols,
3. the need for security while using IP,
4. the ability of remote workers and travelers to access the Internet from almost anywhere, and
5. the relative ease of maintaining IP networks.

So, it is a trade-off. CPE-based IP-VPN solutions provide mobility and security, and they are relatively cheap. Network-based IP-VPNs provide high network performance and quality of service, and service integration and security. A small but growing number of serious IP-VPN providers offer hybrid IP-VPN solutions. Non-mobile, fixed-location users can benefit from the network-based portion of the IP-VPN service, while the traveling sales force or distance, remote office can be supported via a dial-up (or DSL) connection to the corporate VPN. Corporate offices get high reliability, and performance, yet mobile access is also available.

6.2.4. Unified Messaging

Enhanced unified messaging expands access to all messages at any time with whatever device is most convenient. This can include the following:

- Listening to e-mail over the telephone using text-to-speech (TTS) technology.
- Delivering faxes or e-mails to any fax machine.
- Web-based access to messages and mailbox options.
- Global or common directory.
- Single point of administration.
- Mixed-media functionality, such as the ability to attach fax messages to e-mail, annotate and electronically distribute e-mail, or fax messages with voice mail.

People have a variety of communications needs and responsibilities at different times. Mobile and remote workers have a strong need to be aware of messages. However,

being on the road or in a virtual office makes it difficult for them to check e-mail and faxes when they may only have a few minutes at a phone. They may also need to send e-mail or fax messages, but only have access to telephones. Additionally, organizations concerned about productivity believe employees will be more efficient and productive if they are able to check all messages from one interface.

A. Evolution of Voice Mail Functionality

A basic voice mail system records, stores, and plays voice messages. The equipment can be either a stand-alone device or hardware that is integrated within a telephone system. Features permit users to access, reply to, and forward messages; schedule delivery of messages; and tag and edit messages. Integrated systems have message waiting indication via a light on a telephone and/or an alphanumeric display. A ringing telephone can default to a mailbox and telephone callers can leave voice messages in a subscriber's voice mailbox.

Users access messages in their mailbox from telephones or PCs by entering account numbers and personal passwords. With telephone access, information entry and all system commands are performed via the telephone touch-tone keypad or, with speech recognition technology, via simple voice commands. The Telephone User Interface (TUI) of a voice mail system provides the user with voice menu prompting for message management functions.

Call and message management functions of voice mail systems include the following:

- Outdialing to a pager or an external telephone number to notify the user of a newly received message.
- Desktop call screening of inbound calls using recorded caller identification input and/or contact data screen pops.
- Find me, Follow me, and Connect me call screening that puts the caller on hold while the user is contacted and makes the decision to be connected immediately or to request a voice message from the caller.
- Networking of voice messaging systems that are in separate locations.
- Dynamic message delivery checks and returns a receipt for confirmation of message delivery.
- Outbound call initiation (call return) from the voice mail server during message retrieval.
- Personal mailbox administration, including prioritization of inbound messages.
- Sender cancellation or modification of undelivered messages.

- Fax applications, including the routing of a fax transmission that has been answered by the voice mail system to a direct inward dialing (DID) telephone number; treating it as a voice message, with or without an attached voice message (compound message); and fax overflow for busy fax machines. Many voice mail providers have integrated third-party fax servers to accommodate inbound and outbound fax traffic with voice mailboxes.

B. Unifying Message Management Functions of E-mail and Voice Mail

E-mail systems and voice mail systems have been evolving toward unified user interfaces both at the desktop GUI and the telephone interface. Although voice mail system providers made early attempts at unified messaging for legacy mainframe e-mail systems, the results were extremely limited until LAN-based client/server e-mail had established itself.

Three approaches for unified message management have been implemented by the voice mail industry. These are as follows:

- Integrated Servers, Single Message Store - This approach is based on the integration of the e-mail server and the voice mail server and sharing a common multimedia message store (universal mailboxes) and directory. This avoids administration of two message databases and address directories and facilitates forwarding multimedia messages (attachments). Most of the voice mail systems adopting this approach rely upon the multimedia storage, directory, and user client software of those e-mail servers that provide storage and directory functions for voice mail extensions, such as Microsoft Exchange, Lotus Notes, and Novell GroupWise.
- Integrated Servers, Dual Message Store - This is an alternative approach that uses separate multimedia messaging stores and directories for the e-mail and voice mail systems. The e-mail and voice mail servers, however, have to synchronize message and user status information to keep message status up to date and consistent between the two servers and between desktop and telephone modes of retrieval. This synchronization is inherently problematic.
- Client Integration - A simpler, but less effective, approach to unified desktop message management is based upon a software client that separately, but concurrently, accesses individual e-mail and voice mail servers and their separate message stores whenever the user checks for messages. This approach does not permit the voice mail system to provide immediate telephone-based notification and delivery of important e-mail messages, as it would for voice mail messages. However, it does allow legacy client/server e-mail systems to interoperate with voice mail systems. Note: The voice mail server can still function as a client of the e-mail server for telephone retrieval of e-mail messages.

Another level of unified messaging includes cross-media messaging, which covers the retrieval of e-mail and fax text messages by telephone, as well as the ability to exchange voice messages and multimedia information as e-mail attachments. Voice e-mail enables traditional telephone voice mail messages and desktop e-mail-based voice

messages to be sent to unified mailboxes for retrieval, either by wireline or wireless telephone or by multimedia screen devices (PC, laptop, or personal Internet appliance). The two primary interfaces involved with cross-media messaging are the screen-based GUI and the voice-based TUI, which can employ TTS and Automatic Speech Recognition (ASR) technologies. TTS is very important and necessary for retrieving e-mail in voice, while ASR is primarily needed for hands-free, eyes-free situations.

C. Standards

Noted below are some of the more relevant and evolving standards that relate to messaging and convergence:

- *AMIS* - Audio Messaging Interchange Specification is typically used to support analog networking between dissimilar voice mail systems. DTMF tones convey control information and transmission of a voice message in analog form.
- *H.323* - ITU standards specifications are for real-time multimedia communications over packet-based networks.
- *IMAP* - Internet Messaging Access Protocol is for user access over the WAN to Internet mail servers for managing mail folders, scanning message headers, and downloading messages.
- *LDAP* - Lightweight Directory Access Protocol enables client software interaction with a directory service over a TCP connection.
- *MAPI* - This is a messaging API for messaging clients to interact with messaging servers.
- *MIME* - Multipurpose Internet Mail Extension protocol enables the transmission of mixed media data files across TCP/IP networks and is an extension of SMTP. Mime provides the power for document exchange, as well as cross-media messaging.
- *POP3* - This is the Post Office Protocol for e-mail servers on the Internet and is used to access e-mail for downloading. This protocol is key to basic text e-mail exchange.
- *SMTP* - Simple Mail Transfer Protocol is an application-level protocol that operates over TCP/IP for exchanging e-mail text messages between devices supporting Message Handling Service (MHS).
- *TAPI 3.0* - Developed by Microsoft for its Windows PC environment, the Telephony Application Interface enables Windows applications to control key telephony functions common to most telephone systems. Such control, originally provided for direct first-party call control for individual desktop PCs, has been expanded to allow server-oriented, third party call control for call-routing applications. TAPI is important to the industry acceptance of Windows NT as a platform for the convergence of PSTN telephony and IP telephony.
- *VPIM* - Voice Profile for Internet Messaging is a protocol that is being proposed to allow different voice mail systems to automatically exchange voice messages over the Internet.

6.2.5. Presence

If IP communications services are not better, cheaper, or more available, why purchase them?

The answer can be found by looking to the mobile telephony market. Mobile phones are more expensive to buy and operate, the voice quality is poor, reliability is low, and features are only just coming to par with landline. Why then, has it seen so much success? Because it brought significant added value – enough so that consumers were willing to sacrifice emulation of the landline experience for it.

We believe the same is true for IP-based interactive communications (IPIC). To be successful IP communications must present significant new added value. What is this added value? It is the way in which other Internet applications, such as web, email, chat, Presence, Instant Messaging (IM), on-line games and e-commerce can be blended with voice to provide entirely new classes of services and features not available with circuit switched telephony. This added value is changing the way in which communications services are perceived and used – making them more like a web experience than a PSTN experience.

There are strong signals in the industry that validate this claim on added value. This can be observed, for example, in the tremendous growth of buddy lists and other Presence services on the Internet. To date, these services have been used almost exclusively to communicate on-line status in order to facilitate delivery of Instant Messages. However, a much broader notion of Presence, which encompasses all forms of communications, can significantly enrich the IP-based interactive communications (IPIC) experience.

That service offering will come with an open Presence standard, one that is openly embraced. Presence is the key service enabler: as will be described later, Presence takes the current “follow-me” aspect of IP communications to the next level.

Presence is an Integral Part of Enhanced Services

Presence has the potential to communicate not just on-line status, but also:

- Physical location – e.g. “at the office” or “at home”
- Call state – e.g. “ready to accept communications”; “on another call”; this might even include the identity of the other party
- Willingness to accept communication – e.g. “available”, “in a meeting”, etc. This might depend on who is asking.

- The preferred medium – e.g. video, voice, IM, email, etc.

As has been discussed, the compelling value proposition for IPIC lies in new kinds of services. Many of these services are based on Presence.

For example:

"Hop On" – to an Existing Call

If a Presence subscriber could see that one of her subscribees was involved in a telephone conversation, she could request permission to join the call. If permission were granted, it would immediately become a conference call.

Instant Conferencing

If a Presence subscriber could see that all constituents in a potential conversation were on-line, she could immediately invoke a conference call.

Subscribing to Mobile Phone State

If a Presence subscriber was trying to contact a subscribee in transit, he could be immediately notified when the mobile phone was switched on. A combination of this and traditional "call camp-on" would automatically dial the mobile phone as it came on-line. This would supercede mobile phone voice mail which may, or may not, generate a returned call. An enhancement of this service would be to communicate the location of the cell phone user, possibly by polling the identity of the base station to which the mobile phone was connected.

Location-based Screening

Location-based screening would enhance traditional call screening, which is usually based on time of day or calling party, to allow it to depend on the location of the subscriber. For example, a subscriber may purchase a service where only select individuals can call him if he is at home, while anyone can call him when he is at work.

"*Intercom*" *Telephony*

For many years, the international investment community has made use of live connections between trading desks, often on different continents. One example of this was "Hoot 'n Holler", e.g. you pressed a buzzer to alert the other end (Hoot) and then you started talking (Holler). Nextel has implemented this in the mobile domain with their Direct Connect service. However, what is required to make this particularly compelling is some indication that the other party is present and willing to accept communications and in what medium.

Checking Communications State

If Presence becomes a ubiquitous service, subscribers would tend to check the communications state before establishing communications with anyone whether they normally subscribed to them or not. Knowing in advance what the person's communications state was and the methods of communication for which they were available could alleviate the irritation of constant busy signals and "voice mail tag".

Subscribing to the Complete Call State of Another

Full use of the potential capabilities of Presence would be a very powerful mechanism. One implementation would be for an executive assistant to be fully apprised of his boss' call state. This would allow him to interact with her as if she were in the office, even while she was on the road.

The flexibility and power of software-delivered telephony is starting to fulfill the potential of IPIC by providing a platform for an infinite array of services. Our example Presence services are just the beginning. The only limitations are the imagination and ingenuity of the service providers. And the lack of an open Presence standard.

6.2.6. Content Distribution and Internetwork Caching

There are two basic types of network caches, each performs the same function: replicating popular content from originating servers so that it can be accessed more quickly by the user. Prepackaged cache appliances are the first type, bundling hardware and software together. Today these appliances are installed in the carrier or enterprise network, but vendors are already talking about diminutive versions of these appliances for the customer premises, loaded with some 20Gbytes of memory. The second type is caching software, which requires users to purchase the necessary server hardware and install the software themselves. The primary benefit of both types of caching is better performance. Moving content closer to the user reduces the number of router hops it has to traverse, so performance improves. A local cache server may be less loaded and better designed for performance than the originating content server. These factors cut page-load times by one-half or one-third.

There are other critical benefits. Major events such as the release of the Starr Report or an Internet broadcast of a Football WC Final cause huge traffic surges. Masses of users request a particular Web page, which swamps the server. Growing broadband Internet access will only exacerbate this problem. However, replicating the content to other

servers or caches distributes the traffic load around the Internet, alleviating the server bottleneck. ISPs also win with caching and content delivery by keeping more traffic local, using less core bandwidth. Finally, caches can be deployed locally to provide better oversubscription ratios. This is particularly true of cable modem and DSL service providers, where the head-end links can easily be swamped if all users start downloading at once. Cache servers offload traffic, effectively enlarging the upstream pipe and preventing local saturation.

The key then is to shorten the distance between the user and the commonly requested content - not a particularly new science. PC design has long used fast memory caches to feed high-speed processors. Similarly, Web browsers use local-disk caching to improve local Web performance. This is why jumping back to a previous Web page gives the impression of much faster performance.

Network caches work on the same basic principles, but on a much grander scale. Unlike a private cache in the Web browser, however, the network caches select content based on the activity of hundreds and thousands of users. Cache servers learn these patterns by intercepting user requests in one of two ways: transparent cache servers or proxy cache servers. Transparent cache servers sift through all traffic and hence require no modification to the end client. With proxy cache servers, network managers configure users' browsers to direct content requests directly to the cache. The proxy cache server then requests the content on behalf of the user. On the one hand, this lets network managers restrict the sites a user can visit. At the same time, though, this approach is obviously far more complicated, as it requires configuring each client. What's more, if the proxy cache fails, users cannot access the Web.

Content delivery

Content delivery is in effect another approach for keeping content fresh. While caching waits for customers to request information, content delivery lets companies proactively push the information into caches close to the user.

Here is how content delivery works. First, content providers designate content to be replicated to network caches. Here, the network manager either provides a list of elements to be replicated by the content delivery provider or uses tools that troll through the site, marking the elements to be replicated. The content provider then replicates those elements around the globe. Content consumers are then directed to the nearest server through DNS. Geographic penetration of these caches is critical to the success of the

content delivery system. The more caches deployed, the greater the likelihood that users will be able to retrieve a document or element from a cache located geographically nearby. For example, Akamai claims to have 4,200 caches across 50 countries around the globe.

Many content distribution systems are also promulgating their own APIs to encourage developers to create network-aware software. Like the video server pushing a preselected movie into the user's cache, or a main office distributing training videos into the caches of the field offices, these applications can be more intelligent about how they utilize network resources. The result is a content distribution system that begins to look like an interactive and predictive television system. Users can request certain types of content through a Web browser and have it loaded into the cache. Even when users do not request the content, the different elements of the delivery network can work together to anticipate user needs. The cache server can monitor user preferences and send feedback to the content provider. The service can then issue the request to the content server to distribute the most appropriate material at the optimal part of the day.

6.3. CATV – a multiservice platform

Sándor Stefler author

Ferenc S. Tóth reviewer

CATV networks are operating almost everywhere as broadband broadcast programme distribution networks, based on the technology of the eighties. The overwhelming use of coaxial cables for the physical connections and the cascade of 20-50 amplifiers in the long trunk lines are typical. Up-to-date networks however, are using hybrid coaxial and optical fibers (HFC) architecture; these are more sophisticated and can be taken as the pioneers of the much citied future access networks, of the FTTC (fiber to the curb). HFC is a good compromise between the future-proof ness (high bandwidth/transmission speed) and the system costs per subscriber. The major components of such a system are the following:

- headend station
- signal distribution center
- optical node
- network interface

The **headend** processes the analogue and digital signals received from terrestrial stations, satellites and local sources, demodulates amplifies and combines them, makes some I/O transformation then the resulted signal-complex is fed to the network. In the case of digital interactive services here, in the headend are situated the different content (video, mail, Internet, data) servers and other components needed to access the services available. Such elements are e.g. the routers, switches, gateways, etc. For conditional access to some services the different modulations, frequency transformations and/or encoding is realized also here.

The signal **distribution center** transmits the analogue and digital signals from the headend station to the optical nodes in a point to multipoint architecture, using optical fibers. In some minor CATV system the signal distribution center can be missing and the headend is directly connected to the optical nodes. In bigger systems however the signals of digital and local services are taken from telecommunication networks (PDH, SDH, ATM, MAN, WAN) directly to the distribution center, and from here to the trunk network.

In the **optical nodes (ONUs)** the optical signals are transformed into electrical ones and this is fed to the coaxial network. In this network layer (in case of HFC systems) only a few (e.g. two or three) amplifiers are cascaded, which in comparison with traditional CATV systems improves the transmission parameters (especially the nonlinear distortions) dramatically. Such an ONU near to the subscriber supports generally 200 to 1000 subscribers with programme and data signals.

Network interface unit (NIU) makes the network termination on the subscriber side. This can be situated indoor or even in the flat, but it is the most important feature of this unit, that it should be secure and addressable in order to offer subscriber-specific services. Sometimes (if the CATV offers telecommunication services too) to satisfy the requirement of uninterrupted power supply for the continuous data transmission is problematic.

A flexible, future-proof network should be prepared for services still not common today (see Figure 6.3.1). Such services might be e.g. the different kinds of multimedia and different tele-services (teleshopping, telebanking, telemedicine, distant learning, etc.).

You can see on the figure that different services require different data-speeds. The main feature of a service is the required bandwidth that enables the necessary bit rate for that particular service.

For the classification of different services these can be ordered into different **QoS (**quality of service) classes. All services can be ranked into transmission resource quality classes. These are:

Constant bit rate (CBR) services. In this class the transmission rate is constant and stable, so the best quality (e.g. the permissible max. delay and jitter) can be satisfied. This QoS is needed for such sophisticated services as video on demand (VOD).

Variable bit rate (VBR) This service class can be characterized with an average bit rate and with a maximal burst-rate. It is suitable for less critical applications, e.g. Internet.

Undefined bit rate (UBR) The maximal available bit rate is given but not granted. Can be used for the calculation of the free resources of the system. Parameters like jitter and latency are not taken into consideration.

Available bit rate (ABR) This is for services, which cannot be ranked into any of the above, mentioned 3 classes.

Service	Transmission speed		Quality		Isochron transmission	
	Downstream	upstream	Downstream	upstream	Downstream	upstream
Cable telephony	64 kbps	64 kbps	CBR	CBR	*	*
Internet-access	64 kbps-2 Mbps	64 kbps	VBR	VBR	***	***
Teleeducation	64 kbps-2 Mbps	64 kbps	CBR/VBR	VBR	**	***
Teleworking	nx64 kbps-8 Mbps	n x 64 kbps	CBR/VBR	CBR/VBR	*	*
NVOD	2-8 Mbps	kbps	CBR	VBR	*	***
VOD	2-8 Mbps	kbps	CBR	VBR	*	*
Video conference	n x 64 kbps-8 Mbps	n x 64 kbps-8 Mbps	CBR	CBR	*	*
Telemetric	kbps		ABR	ABR	***	***

Figure 6.3.1. Transmission parameters of different multimedia services

Different modulation schemes set different requirements concerning the transmission system. The modulation scheme used in CATV systems for transmission of data requires a data rate depending on the application of the data. And this has to be granted at a carrier-to-noise ratio (C/N), which can be realized in the channel. The bit-error-rate (BER) can be utmost $2x10^{-4}$ as the error correction circuits usually cannot correct values worse than this. Higher transmission speeds require better C/N values, and where this cannot be realized, more robust modulation schemes should be used, which allow the signal transmission. in frequency ranges more critical from the view of noises (this range is usually the frequency range of the return channel). The bandwidth-efficiency, which can be aimed at theoretically, cannot be reached in real systems, because of– among others – the limited spectrum available. Real values are only slightly below the theoretically values however. Based on these facts, major manufacturers choose for forward transmission the digital modulation 64 QAM, while for return channels QPSK.

With the examples of cablemodems and cabletelephony in the following we should like to highlite the introduction of tipical, actual multimedia services into CATV networks.

Cable modems and cable telephony

In order to use CATV networks for data transmission and telephony, two-way transmission, i.e. suitable downstream and return channels are required, with bandwidth according to the type of services. In the case these conditions are not met, the network should be upgraded accordingly. The frequency plan of an up-to-date CATV network with 862 MHz bandwidth can be seen on Figure 6.3.2. Today the frequency range of 5-30 MHz is available for the return channel. To satisfy the capacity-requirements of the expected future services this frequency range should be extended at least up to 65 MHz. That means TV band I (47-69 MHz) cannot be used for broadcasting TV programme. Therefore some measures should be taken in order to give up TV band I. for broadcasting. Of course

this can be done only after suitable agreement between the regulation authorities and broadcasters.

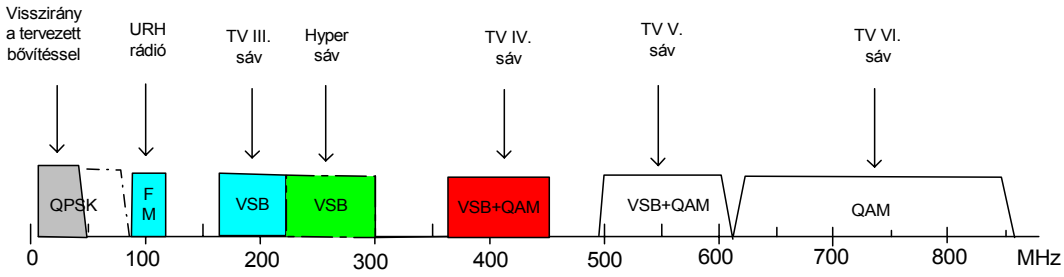


Figure 6.3.2. Frequency plan of up-to-date CATV networks

6.3.1. Utilization of cable modems

Cable modems can be used in two-way CATV systems for high-speed point-to-multipoint data transmission, or LAN-to-LAN connection in full duplex operation mode. The most common application today is the high-speed Internet access for the CATV subscribers. It is in commercial application in Hungary too. Cable modems can be used in coax-only networks as well in up-to-date HFC networks. These modems compared to their analogue counterparts can be characterized by extended functions. They have usually the capability of a router with added TV-tuner features and network management and diagnostic features. Some cable modems have conditional access and authorization functionality too. They use in the USA a 6 MHz wide band, while in Europe usually 8 MHz. and are frequency agile. The return path is different in the systems, but typically between 5 and 120 MHz. They search for a clear part of the spectrum and tune them self to this region.

A full data transmission system over CATV consists of the following main elements:

- cable modem,
- headend transmitter/receiver
- data control system (special cable modem, with interfaces to the routers, switches, bridges).

On the subscriber side the cable modem has an RF connector (typically F-type) towards the CATV network and an Ethernet connector (RJ 45) towards the computer or router (in the case of SOHO application). The data rate on a 6 MHz RF channel is up to 30 Mbps in the downstream and up to 10 Mbps in the upstream.

According to the number of allowable RF channels several 6 MHz channels can be operated in parallel and combine them. Most cable modems can be connected to a network management system, usually SNMP-based.

The headend (node) transceiver is a frequency transposer unit, which transforms the data coming from the upstream into the downstream. To achieve high reliability this transceiver is often supplied by a redundant power supply and with an automatic switchover.

The node data controller is the gateway between the CATV network and the telecommunication network. Technically this unit is a special cable modem with added router features.

A reasonable solution for extension an existing network is to integrate it with a node remodulator. By processing the IF signals of the node transceiver it separates the downstream and upstream signals and converts them into baseband. The node remodulator at the same time corrects the timing of the signals and removes the disturbing signals from the upstream and produces regenerated output with defined level at the headend. The node remodulator can be useful on sites where the return channel is seriously disturbed by unwanted signals.

6.3.2. Cable telephony

By the liberalisation of the PSTN services the last segment of the telecommunications will be opened too for the competitions. This will offer CATV operators new challenges. If they can afford this, they may hope to extend their service packages and to step forward to become unlimited telecommunication service providers. Especially cable telephony however will raise the requirements towards the infrastucture significantly.

From engineering point of view cable telephony can be realized by two different way: IP (Internet-protocol) based or using a special type of PABX. Today this last one is more common however it is clear that the IP-solution is more sophisticated but is still expensive.

An important element of the system architecture is the distribution point and the coaxial network termination (NIU) at the subscriber. The distribution point represents the gateway between the CATV and the public telephone network using switched or leased lines. In this the following functions are realized:

- modulation/demodulation of the data and speech signals,

- processing and transformation of the signalization,
- concentration of the data- and speech-traffic,
- providing an interface between the switched an leased lines,
- providing an interface towards the network management system.

The NIU realizes the transition from the HFC network towards the subscribers. The main functions of this NIU are the following:

- modulation/demodulation of the data and speech signals,
- converting the RF signals into baseband,
- monitoring and measuring the bit-error ratio,
- looping back to track errors in the transmission,
- looping the RF input signal to other terminal units, like TV receiver, tape recorder, etc.

Only very seldom occurs that the btelephone exchange colocates with the CATV headend, therefore the interfaces at the distribution node have to enable connections toward the telephone exchange, using standardised transmission systems (PDH, SDH, ATM). Similarly occurs the connection between the distribution node and remote exchange using a concentrating or non-concentrating 2 Mbps interface (V.5.x)

In order to successfully compete the new service providers with the existing ones, at least similar services should be offered. That means they should prepare themselves to the following services:

- POTS (plain old telephone system)
- ISDN basic access,
- ISDN primer multiplex connection,
- leased lines using n x 64 kbps channels (V.11, X.21) and/or 2 Mbps (G.703/704 synchron)

In order to keep the expenses per subsriber at minimal level, it is useful for the NIU to offer POTS and ISDN leased lines. Such a way the subscriber can choose the service variants mostly conforms his needs. An important feature is to maintain lifeline services i.e. to keep the system in operational state even in the case when the mains voltage is interrupted. As most existing CATV networks do not enable the remote powering the NIUs, these should be powered from the local AC mains. In this case however the telephone service will be interrupted together with the mains break down, similarly as is the case with cordless telephones. There are however subscribers who want (or must) telephone in these conditions too (so called lifeline services). Therefore it is neceessary for the

subscribers of the cable telephony to solve the uninterruptible powering their NIUs (eg. using batteries).

6.3.3. Network management

Now we can see that new services require much higher capacity and reliability of the CATV network infrastructure than it is common with the traditional, pure programme distribution networks. In order to comply with the requirements of telephony, CATV operators must have more extensive knowledge of the actual state of their network, i.e a network management system has to be established with the following main features:

- supervisoring of possibly all network units,
- automatic switch-over to the standby system,
- continous monitoring of critical system parameters,
- displaying all previously mentioned results on an easy-to-understand graphical terminal.

The so called “element management” system is communicating with many differrent units in the network, in order to evaluate the system informations. In a complex network many, different element management system may exist, each of them formed to the specific needs of particular units. Such units can be e.g. amplifiers, power supplies or even complete cable modems. The element manager checks the following:

- parameters, signal levels, measurement results,
- alarm signals,
- the traffic conditions (free capacity) of the system,
- bandwidth- and traffic management,
- system checks,
- security management,
- spectrum analysis.

The element manager operates under a higher level so called “domain manager” which summarizes all the information supplied by all element manager systems. Data exchange is made using standardized interfaces and protocols (e.g. SNMP), and the results are forwarded in a compressed form to the highest level of the network management system. The “service management” layer gets tremendous quantity of data. On the highest level is the “business management” system which is connected with the subsriber management system (SMS) too. If well planned this complex network

management system will help to run the multiservice CATV system in good (reliable) efficient and economical condition.

6.3.4. Migration from the CATV to the interactive digital networks.

The value of a broadband cable network is directly influenced by its profit-generating capability. This has serious consequences on the network planning or on the eventual modification of an existing one. It is because the network architecture determines the services it can offer in the following 3 to 4 years. Therefore before making decisions the answers on the following basic, but very important questions should be evaluated:

- what is the relation of the network to the Internet,
- how return path communication is handled,
- what security can be granted.

At the dawn of the digital era the convergence and integration of networks was forecasted. And this came more or less true. In our case that means, that the connection between CATV networks and telecommunication networks is now under way. We can observe the forming of the network of connected networks, in order to offer TV programming, telephone, data transmission and interactive multimedia (above all Internet) services for individuals and enterprises too.

From such a viewpoint the CATV

- cannot stay anymore one-way programme distribution network, but should be transformed into broadband interactive network. According to our present-day knowledge the HFC architecture is the one, which has the best performance/price ratio for that purpose.
- should take into account the challenge of telecommunications and media sectors.

In order to realize the transformation of CATV into interactive broadband network the following decisions should be made:

- evaluation of the market background of services to be offered,
- the timing of the necessary steps,
- defining the schedule for the technical tasks (to change from analogue to digital, from coax to fiber, from one-way transmission to interactive, from local functioning to at least regional, from simple operation to managed networks, from free distribution-distribution to conditional access and secure communication)

From the above-mentioned procedure some requirements can be derived for the digital broadband networks:

- wide bandwidth

- real time, scalable return channel
- secure authorizing with public keys,
- end-to-end IP capability,
- sophisticated network management
- sophisticated, standardized interfaces and terminals (set-top-boxes)

While it is inevitable that the networks are those which define the possible choice of services, but the importance of set-top-boxes should be emphasized, because they play the key role in accessing, convenience and security of services. Up-to-date set-top-boxes are network- and purpose-oriented computers, with a lot of telecommunications, CA- and media-interfaces, with sophisticated, IP-oriented operating system (e.g. OpenTV, MediaHighway, or Media Home Platform), much memory, and sophisticated graphical capabilities. It is very important that these boxes should manage efficiently the return path communication and should enable an easy migration toward new features. And they have to be very stable and user-friendly. Therefore set-top-boxes are critical for the success of digital services, but they can fulfill the tasks only in full harmony with the whole network.

Summary

Modern CATV networks offer high-power, efficient and economic infrastructure for broadband services. Future-proof networks should be two-way broadband networks with flexible headend and subscriber terminal, in order to enable several high-speed interactive multimedia services. The high-speed Internet-access and cable telephony are messengers for a series of profit-generating, value-added telecommunication services on the CATV networks.

After the liberalization the operator of CATV networks can be an alternative telecommunication service provider, if it is ready for that. That is in short time the pure programme distribution networks should be transformed into full-scale (FSN) networks. Of course this procedure requires the modernization and equipping them with new components. And this is already the era of interactive digital television.

6.4. Telecommunication terminals

László Binder author

Ervin Kovács reviewer

6.4.1. Classification.

This chapter may appear as the odd one out dealing with the telecommunication services (or those accessible through telecommunication) from the aspect of the final user (the human being) between the strictly technical (or informatic) ones.

This approach seems nevertheless well based given that the number of those participating in the development or manufacture of the equipment and services, writing of programs or running the services – those who will represent the majority of the readers of this book – are largely superseded by those using the services. (This shall be evident only by mentioning that among the 10 million Hungarians roughly 4 million use landline telephone and approximately the same number of mobile phones are used too, statistically every household disposes of a TV set).

There are different groups on the user side too, the needs of which - but also the capabilities to absorb services - are considerably different. Already the Telecommunication Law treats the disabled in a pronounced manner ordering and promoting the services specially applicable by them. It is not necessary however to go as far as that. It is sufficient to appreciate the young people growing up together with the internet and at the same time the elderly generation living mostly at the country-side, (afraid of everything presenting a technical character) nevertheless being the target group of a number of informatic applications (presented in the relevant chapter).

The technicians – and the internet users presenting often similar needs – are rather often served by traditional computer equipment, occasionally optimised to their application. The ergonomic issues treated below more in detail are relevant in these cases as well.

With respect of the subject of this chapter we suggest a definition that differentiates of the general meaning of the word. We deliberately limit (or extend?) the meaning of the word of telecommunication (or „telematic”) terminal to such equipment and software linked to it that enables the human user of any service to use it through telecommunication

system(s). These terminals deserve high attention as they permit the comfortable everyday access of the public to the new services therefor constitute a key element of the widespread usage of them. Their role may be compared to that of the saddle: riding a horse is possible without it and some specialists do it regularly. But using the horse as a transport tool by ordinary people and for long distances would be cumbersome if not impossible without.

This chapter deals more in detail of the equipment providing the interface between techniques and user - other than the above mentioned specialist - so to say the “general user”. As the number of the different applications for them is already today huge and grows daily it is very difficult to establish closed categories. A foothold may be found on the ergonomic bases characterising steadily the human being in spite of its generally cited adaptability and development.

On the outbound side of the technical-human interface (perception) meaning the communication in direction from the technical side to the human one the vision and the audition are playing the most important role.

On the visual field the following may – in function of the given application – be of importance:

- The full size of the image, its brightness, contrast, angle of visibility,
- The picture resolution and colour depth,
- The timely validity of the image, need of hard copy

The channelling of auditive information seems to have less criteria, as:

- Sound quality (frequency range, dynamics)
- Mono-stereo
- In case of speech the language(s)
- Continuity, latency

Some particular solutions may be applied to solve specific demands e.g. text presenting to the blind either by using tactile (Braille) sensors or reading programs.

On the input side that means transmitting from the human to the technique the human capabilities dominate as well. The sizes and the physical characteristics of the traditional input tools as keyboards or even tactile screens are defined by the measures and possibilities of the human hand and fingers, sometimes depending of the category of the user too. In case of stable allocated equipment the possibility and way of remote controlling may be important. Voice input or control is emerging and applied already in

everyday usage (like dial triggering in mobile phones). Controls using eye movement or bio-current exist now in military systems but my become generally applicable even before this book gets obsolete – event probably arriving soon.

The telematic terminals dispose of additional interfaces compared to the general usage informatics-telecommunicaton equipment. This is due to application-oriented and security reasons. Some typical examples are magnetic stripe or chip card reader-writers, bar code readers, scanners, bill- or ticket printers, video- audio connectors.

An other aspect of regrouping terminals is there place and type of usage. We have to distinguish mobile tools from those firmly installed, serving one or more users at the same time. The realisation of the terminals may relay here on the ergonomic factor as well. Whereas the stable terminals (or those “hardly moveable in the reality but marketed like mobiles” as defined by the late professor Richard Kolos) shall provide maximum (sensible and valuable) quality of picture and voice reproduction the really mobile ones (like mobile-phones, PDA) may be limited to the readable, understandable level (though the higher quality contributes to the competitivity of the product), in case the size and weight corresponds to the needed everyday usage criteria. This does not obligatorily represent minimum size or weight, as for example the size of the human fingertip limits the miniaturisation of the securely usable keyboard before the technical possibility.

Finally the set of applications and the tools (equipment and software, including telecom technology) linked to them is limited by the objective and subjective human demands. As any new technology asks for efforts (financial, like investment on the service provider side, usage cost by or on behalf of the user and time and work for preparation on both sides) the opportunity cost of the application (or service provision seen from the other side) has to be accounted for. This should be compared to the advantages, sometimes difficult to calculate or measure (like being contactable or able to ask help in emergency by mobile phone, ease of gathering information about events at a kiosk, possibility to co-ordinate a TV movie with the available free time slot). In addition to all of that we shall not forget that already in the 20th century as today demands are not getting satisfied by the solutions but rather generated by the marketing based on new technological capabilities. (And this is not specific for the telecommunication.) Therefor the particular forms and timing of appearance of the equipment described in this chapter depend non less on the interest of the developer-manufacturer companies as on the described technical and human factors.

6.4.2. Examples

With the aim to concretise the above considerations not usual in a technical handbook we shall provide some examples below. There choice is not arbitrary, it is driven by the intention to show some rather typical solutions present at different fields of life together with the equipment linked to them. At the same time they are rather well separable in the categories (not detailed above because it is nearly impossible to define the most relevant among the uncountable criteria).

The first example is a typical fixed home appliance accomplishing high level services. The – naturally moving – pictures are aimed to be conform to the high definition TV standard, the voice to be stereo and multi-lingual.

The second example addresses a large population in the frame of the information society services providing a tool for the introduction of administrative, information, education and financial services avoiding at the same time the employment of a large number of specially or at least highly educated personnel. It combines the business interests of the service providers with that of the public administration but provides the personal benefits of the users (more and better information accessible with less time and financial efforts) at the same time.

The third and last example is the personal digital assistant replacing efficiently and without paper the everyday tools (and what's important, the functions they are used for) of the businessman today and those of the student or housewife tomorrow.

It was indicated in the introduction that this chapter shall not deal with the typical internet users as their terminals are basically standard or a little bit tuned PC-s. We shall mention however that all three of our examples include the possibility of running internet applications as one may not imagine a telematic terminal lacking this feature.

The following descriptions may reflect particular manufacturers products (as there is no theoretical equipment) but are not intended to use them as reference or standard. The mentioned equipment and application are not detailed either as other chapters do so in a more exhausting manner, the intention is to demonstrate the link between the equipment, application and the user.

6.4.2.1. Integrated home multimedia terminal

The today's household is equipped by at least one radio receiver and a TV set. hi-fi equipment being optional and home computers are only making their entry into the list. However the need for connectivity to both cable (or satellite) multiple TV program source as to Internet emerge parallel. Different solutions are either offered already or predicted for the very near future. They will need combined terminals in all cases, but the terminal's features will differ in function of the local demand. Let us see the possibilities together with the influencing factors:

a.) Independent audio-visual centre and home computer (or Internet terminal) This solution is applicable where the parallel usage of both equipment is possible (independent rooms) and needed (different usage habits of family members). Even in such cases combined telecom terminals will most probably be used as the combined connectivity (cable or satellite) certainly advances the separate telephone and receiver solution not only in economic but also in performance aspects. In this case a connection/hub box provides the telecom link for all listed units.

b.) TV set performing integrated Internet terminal function. The configuration is similar to a traditional TV set with decoding set-top box but this unit is providing a complete terminal's features too. The architecture is naturally based on a PC and incorporates all functions usually provided in the TV-video environment (i.e. TV set with mono- or stereo sound, video recorder, DVD player, satellite receiver/antenna driver) as well as the interactive Internet service centre permitting browsing, e-mail etc. usage. The functions are realised through the PC, in consequence their performance exceeds traditional levels and their combination permits new usages. Some examples: the recorder uses a high capacity hard drive instead of tape, thus the search, replay is practically immediate. The programming is using clear-text guided written on-screen communication through alpha-numeric keyboard, pre-defined passes or cumbersome button-choice menu handling disappears. On the other hand, Internet browsing happens through high throughput broadband channels, waiting

time (and possible PSTN charges and conflicts with traditional telephone usage are eliminated). In addition to the increasing quality of traditional functions new ones appear, like video on demand or pay-TV with the possibility to watch a pre-recorded program at a chosen time (and pay only if you really ran it, not if you recorded it for the eventual need but discarded later on). Features of digital broadcast TV like multi-lingual voice etc. become evident. It is visible that this type of terminal – though not cheap at the moment of this report – will become very widespread. One of the existing configurations is shown on the picture below. The eminent distinguishing feature of it is that it looks like any other home video equipment. (The keyboard shown is optional for frequent Internet users, but the remote controller contains a full miniature one if you open it!)

Technical data sheet of a typical home telecommunication terminal:

TV set:	min. 69 cm 9/16 format
	100Hz "flat" CRT or TFT/plasma flat screen
	NICAM stereo sound
	2*SCART + audio/video RCA input/output plugs
	240V AC
Set-top box:	PC controlled,
	Windows or Linux operation system,
	resident application software
	128 MB RAM
	min. 20 GB, typically 40 GB hard disk
	DVD player
	smart card based access control
	infrared (or Bluetouth) keyboard (standard size + miniature)
Communication	unit linked to broadband asymmetric telecom service:
	ISDN (min. 128 Kbps) or
	ADSL (24 Kbps uplink/min. 400 Kbps downlink) terrestrial or
	broadband satellite (min. 38 Kbps uplink/2-8Mbps downlink)

6.4.2.2. Public information terminal

As a contrast to the home telecommunication terminals the public information terminal is integrating multi-functional content with robust casing, compact self-contained design and particularly easy operating procedures. Its typical user is non-professional, needing access to public or functional/professional services like municipal administration, tourist/travel information etc. services without taking advantage of the personal communication of a specialised agent over the counter. The terminal has to be able to enable access to free of charge information as well as to such provided against money. Therefor the terminal shall be connected to banking (credit or debit card) networks or accept local payment by cash or electronic purse. This capability extends the application field to the sale of tickets, mobile-phone prepayment etc. but also to money transfer type (payment of electricity/gas bills, e-commerce) usages. Such terminals – called also "kiosks" differ from traditional ATM by having no (or limited) cash storage, thus its

manufacturing and installation cost is significantly lower, therefor more affordable both for the application provider and through it to the end user.

Due to the above described usage the equipment hardware-wise has to be solid and simple at the same time, permitting that non-trained (computer-illiterate) persons understand and accept their usage. They typically contain touch-screen displays with additional keyboards (mostly numeric-only), card readers and hard-copy printers.

Typical public information terminal specification sheet:

Size and weight	1700*600*500 mm	100 kg
Mains	240 V AC 600W	UPS 15 min
Display	12”-14” LCD VGA/SVGA, min 16 bits colour	
	touchscreen resolution min 4K*4Kpoints	
	anti-vandal armoured glass	
Keyboard	numeric/alphanumeric,	
	anti-vandal, metallic	
	DES security protection	
Card reader	magnetic stripe reading	
	optional smart card reading	
Printer	receipt printer (thermal, 40 char/line, thermal, 8dpmm	
	A4 printer (sheet or thermal, 300/600dpi)	
	Optional scanner feature on A4 printer	
Audio/video	optional phone (VoIP or ISDN)	
	optional camera and video-conference feature	
	optional stereo multimedia	
Computer	PC with Pentium III, HDD, CD-ROM	
	Windows, Linux	
Applications (some typical examples) :		
	Internet browsing	
	Document downloading/rescanning	
	Data-base consultation	
	Interactive information system	
	On-line reservation/ticket sale	

6.4.2.3. Integrated mobile personal terminal

Some fifteen years ago the European businessman visiting the USA was shocked about the widespread usage of mobile phones in public places. Ten years ago having one here in Europe was a status symbol, five years ago the parents began to attach one to their kids making them “connectable” and today in numerous countries the penetration of mobile phones exceed that of the terrestrial. Listening to one anywhere (except particular places or events) and speaking to one (if not too loudly) is no more remarkable.

Some fifteen years ago the first really usable lightweight laptop PC appeared permitting the businessman to have his data available anywhere. Ten years ago the “pen-top” PC appeared abandoning the keyboard, using the flat screen as tactile input device instead. Five years ago the combination of a mobile phone with a “manager calculator” made its first attempt to replace a notebook (weather paper or laptop computer based) and

6.4.2. ábra. Nyilvános multimédia terminál

is still struggling to get accepted. Some began to combine those tools and the PDA (Personal Digital Assistant) was born. Today we may witness the arrival of this latter at the level the mobile phone had ten years ago, that means that not only American businessman but the public uses them in the everyday life. What's more those using them become dependent on them, again as the are on mobile phones.

What are those small devices offering to enable that?

First of all they replaced some traditional tools (e.g. agenda) in a most "trendy" way,

showing the same effect as the mobile phones did in the initial period: it was a status symbol having and occasionally using a relatively expensive “modern” device. The time passing the data accumulated in them became vitally important as there was no easy and fast way to reload them into an other storage medium. In parallel the manufacturers (and their hardware and software developing application partners) realised that some add-on units enlarging the original application field are needed. The effect of positive feed-back appeared and permitted that the equipment today is

Affordable

Widely usable

comfortably usable

no more a status symbol but rather a useful appliance

The industry standard(s) emerging the multitude of optional devices made of the “metallic agenda” a real telecommunication terminal as well without needing it to be linked constantly to a network. It is the ease of connection to different types of communication channels (infrared, Bluetooth, traditional PSTN plug, GSM, or even satellite) that makes a mobile, quasi-universal terminal and a quasi-universal personal assistant (it’s now that the name is really true). At present the – previously reluctant – user changes to an addict. And maybe – as it happened (first of all on the German-language markets) to the mobile phones becoming “handy-s” - the sophisticated name: PDA will change to “palmobil”. (Suggested by Mr. István Bartolits on occasion of the CEBIT 2001)

Typical integrated mobile terminal specification sheet:

Size and weight	11*8*1,5 cm	110-140 g
Display	B&W or 16 bit colour LCD, backlight , 5,5*5,5 cm	
Memory	4-8 Mbyte RAM, 2-4 Mbyte ROM	
Operating system	PALM-OS®, Windows CE®	
Communication	infrared, Bluetooth, serial, USB ports	
Accessories (functional only):	communication kit (modem or GSM)	
	backup/extension memory	
	keyboard	
Standard applications	address/phone list	
	to-do list	
	notes	
	expense register	
	calculator	
	e-mail	
Optional applications (several hundred, among them):	word processing	
	spreadsheet	
	navigation	
	phone connection (auto-dialling etc.)	
	SMS, WAP	
	e-book (downloadable literature)	
	photo-, video making or reviewing	

Conclusion:

Telecommunication terminals are inevitably entering every field where the human being, the end user of a multitude of services has to be connected to the service provider, independently of the location (home, public places, fixed or mobile) and of the character of the service itself. The devices – though mostly PC based – may be very different but are characterised by the fact that their usage does not need computer knowledge. This is important in spite of the tendency of computer usage getting popular as well. These terminals distinguish of being application oriented rather than technologically determined again in spite of the fact that most of them are permitting the access to different kinds of applications due to their architecture and software.

The above “definition” tries to differentiate these terminals from the multitude of computers naturally used in network environment and functioning as terminal equipment of such complex systems as well as from hardware and software components (like modems) linking local equipment to remote servers or between them but not being directly in contact with the end user, the human.

6.5. Online services, e-commerce

Péter Bakonyi dr., András Ercsényi, authors

György Takács, dr., reviewer

6.5.1. Concepts, definitions

Possibilities for new communication in the form of new services and new products appear in public administration, education, culture and in other spheres of life. The whole process, which started in the United States by now is almost everywhere. The economic system known as *e-commerce* has significant relevance in the countries of the European Union as well.

In Hungary the Telecommunication Act[2] contains the working definition of e-commerce. According to that e-commerce is a type business activity based on the electronic processing and transfer of different types of data, such as text, image or voice. E-commerce includes for example the:

- electronic distribution of products and services,
- electronic transfer of electronic money,
- making out of an electronic bill of freight,
- electronic management of commercial auctions,
- electronic implementation of public procurement,
- activities of an electronic customer service,
- and other activities not listed here.

This definition however is of a very general nature. On this basis business activities through the Internet, shopping through television with videotext or CD-ROM based marketing could also be named as e-commerce.

Today e-commerce has another definition gaining more and more acceptance, according to which e-commerce means a type of business activity based on electronic processing and transfer where the World Wide Web[3] plays a significant role.

According to a different definition, also in use, e-commerce means:

[2] Hungarian Telecommunication Act 2001/XL. – came into force in July 2001.

- the possibility to conduct business through electronic channels,
- new and effective means of product marketing,
- the importance of electronic customer-service and its follow-up,
- the relevant role of the World Wide Web.

Further definitions of e-commerce are the following:

1. E-commerce is the type of selling and purchasing of products and services where the transactions partly or fully take place with the help of electronic channels.

2. Together with the help of e-commerce two or more participants can exchange products, goods, services or contents; or manage financial transactions using electronic tools and technologies.

The notion of e-business is also often being used. In this expression 'business' denotes any possible activity, therefore e-business is a less specific term than e-commerce.

It is important to note that in all the expressions with the prefix 'electronic' – like in electronic-commerce, electronic-procurement, electronic public notary etc. – the adjective refers to a possibility to connect in a network through an electronic medium. It can also be paraphrased that 'e' presupposes distant access. This type of connection can also be indicated by the Greek prefix 'tele' like in word television. Therefore instead of electronic commerce the expression tele-commerce can also be used.

However we will use the expression 'e-commerce' for further reference.

6.5.2. Areas of e-commerce

E-commerce distinguishes among several big areas. The most essential – being the first one to develop – is the information and communication technology based business activity between businesses. According to the American terminology this is called the 'business to business' (B2B) branch.

The new form of retail trade, that is the connection between the small business and the consumer is called 'business to consumer' B2C being the abbreviation.

There is a new branch of electronic commerce between businesses and public administration, which primarily collects and provides information and client service. This is called 'business to administration' (B2A).

---

[3] World Wide Web is the sum total of linked document to be found on the Net. Every document has its own name, which is called URL (Uniform Resource Locator) or simply web-address.

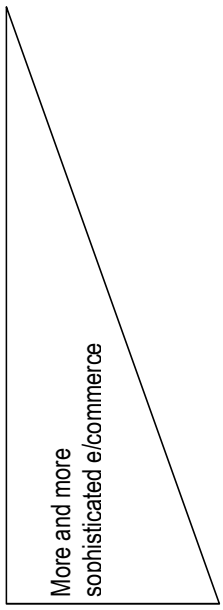
Relations aiming at the exchange of information and communication between citizens as consumers and the public administration are becoming more relevant. An example for this could be the electronic tax return or the electronic filing of a request. This branch is called ‘consumer to administration’ (C2A).

Obviously e-commerce supports the entire business process, beginning with developing the requirements, through marketing and sales and all the other services that they entail. In the next chapter we will show what units, phases the above mentioned business process contains.

All those participating in e-commerce have to be able to collect, search, navigate, filter and forward information. They also need to be able to exchange commercial contracts and other objects relating to e-commerce and access other electronic services as well.

6.5.3. The classification of e-commerce

Phase 1 General Marketing	Presentation and marketing on the Internet. Information about products and services presented on static web-pages.
Phase 2 Special marketing	Presentation and marketing on the Internet. Dynamic information with interactive possibilities.
Phase 3 Contracting (buying/selling)	Presentation marketing and sales on the Internet. Interactive, providing the possibility to order online. Payment information (number of the credit card) can be transferred through the Internet. Typical example: online purchase, online market. Online sales do not contain the payment transaction and the delivery-distribution.
Phase 4 Payment	Presentation marketing and sales on the Internet. Interactive, providing the possibility to order online. Payment transaction on the Internet with the help of Internet payment tools (e.g. smartcard). Online sales contain the payment transaction but not the online distribution.
Phase 5 Delivery-distribution	Presentation marketing, sales and delivery-distribution on the Internet. Interactive, providing the possibility to order online. Payment transaction on the Internet with the help of Internet payment tools (e.g. smartcard). Online sales contain the payment transaction and the online distribution as well. Typical example is the distribution of software.



In the **first phase** the companies participating in e-commerce have a Web presence already but that contains only basic information, which is not up-dated on a regular basis. Interactive access is also limited. In most cases the existing printed forms are up-loaded on the web in a digitalized format.

In the **second phase** the interactive features become more important. A good example is a Web sales catalogue where one can search interactively. This can be supplemented with a real time access to the availability of the required products.

In the **third phase** this is further developed by the possibility to order online. Payment can be made by credit card on the Internet or by cash at delivery or by other means.

In the **fourth developmental phase** the web page offers online payment (e.g. e-cash) but this is not very widespread yet.

Finally, in the **fifth phase** the entire transaction is implemented online, including delivery as well. This presupposes a type of product that can be easily digitalized and transferred. A typical example is the sales of the software.

6.5.4. E-commerce in Europe

Infrastructural indicators in EU countries

In the United States sales by e-commerce in 2000 amounted to almost 1 billion USD. In other developed countries the current share of e-commerce in the GDP is around 1-2% but it is expected to rise to 6% by 2003. The bigger part of European (economic) organisations are well prepared to participate in e-commerce. 62% of the organisations use e-mail and 66% have Internet access. More advanced IT solutions, such as intranet are less widespread (31%) but they are also expected to develop fast.

Regarding the general population 44% of the European citizens have access to a PC at home, but the possibility for online access from home differs from country to country. There are some countries where this indicators is above 30% (the Netherlands and Denmark) but in Spain it is under 10%.

In the ten most developed European countries (90% of the EU's population) there are 48 million e-mail users.

It is not surprising that 95% of the EU population knows about the Internet and 35% have already used it. This ratio is above 50% in the Scandinavian countries. In the above mentioned 10 EU countries there were 54 million Internet users at the end of 1999 and this number will rise to 135 million by the end of 2001.

E-commerce in the EU countries

The most popular e-commerce service currently is the search for information regarding the sales of products and available services (e.g. travel information, hotel booking).

The online purchase that is the online ordering combined with the postal delivery or immediate downloading is also quite widespread. 14% of the population uses such service and by the end of 2001 this number is supposed to double.

Online banking is another fast developing area. By the end of 2001 the number of people using this service will triple, reaching 20% of the population this way.

6.5.5. E-commerce in Hungary

Infrastructural indicators

The situation regarding the domestic private sector is favourable. In the middle of 2000 53% of the companies had access to Internet. By the end of 2001 it is expected that 95% of the bigger companies would have Internet access.

Around half of those companies employing more than 10 people had own their web presence at the end of 2000. This number is expected to rise further in 2001. 10. 14.

Numbers relating to the general population are not that encouraging. PC penetration rate in the population in 1999 was 7.5%. This number is lower than many of the neighbouring countries'. The reason lies with the high import taxes levied upon computers and parts in Hungary.

According to an official survey 10% of the households had access to a PC at the end of 1999. This means 400 000 PCs in absolute numbers.

The Internet penetration rate is also lower than in some neighbouring countries as it stands at 8-9% at the moment. 3-4% of the households have Internet access. The slow rise is caused by the high costs of Internet use, the most expensive in Europe.

E-commerce

The B2B type of e-commerce is relatively well developed and the trade increases every year. According to the estimate of Carnation Consulting the 400 million USD B2B business in 2000 will increase to 1.4 billion USD by 2002.

B2C, that is the e-commerce from business to consumer is rather undeveloped. In 2000 business amounted to only 0.1% of the entire small business trade with 570 000

USD. This branch is also expected to develop dynamically and the yearly business will soon double.

6.5.6. The legal background of e-commerce

E-commerce raises several problems, among which legal issues are of exceptional importance and also of diverse nature. The following issues are especially relevant:

- electronic authentication, identification and encryption,
- general contract conditions from the point of view of e-commerce,
- advertising rights
- copyright.

  We will deal with electronic authentication in more detail.

Electronic authentication, identification and encryption

One of the fundamental bases for the use of e-commerce is the appropriate and reliable identification of participating partners. Very often another, almost as important of a criterion is forgotten: the exchange of commercial messages between the participants has to be unreadable to the third parties. The first set of questions concerns authentication and identification, while the second the territory of encryption.

The basic problem of authentication raises the question whether the communicating partner is really who s/he says to be. This is a very real problem as on the Internet anyone can send an electronic message in the name of someone else. At the same time a programmed server can also illegally 'personify' any enterprise.

Authentication represents a problem because in many cases the legal requirements prescribe 'written' presence, that is apart from having a document, this also has to be signed. Therefore, the validity of electronic contracts, electronic authentication and electronic signature are problems to be solved.

Today not only in the United States and in the EU countries but also in Hungary there are legal regulations relating to this subject. According to the 'legal principles[4]' an electronic document is a declaration or the acceptance of a declaration in the form of an electronic text and as such it is an electronic document at the same time. An electronic document is also a type of data that can be detected through electronic means and it is

[4] First paragraph of the 2001/XXXV. Hungarian Act defines shortly the notion of an electronic document, record etc. We used these definitions.

supplied with an electronic signature. Finally, the electronic signature, which is attached to the document – or is obviously connected to the document –is a type of data that can identify the signing party unambiguously and it can also show whether the original content of the electronic document has been altered or not.

Further on we will show the background for the technical solution of the electronic signature. The electronic signature in use today is based on the so-called 'public key' encrypthion (RSA[5]). According to that everybody has a pair of keys capable of coding. The key-sets relate to one another and it is not easy to produce them as it requires a mathematical apparatus based on a large number of prime numbers. One member of the key-pair is deposited in a place accessible to the larger public[6], this is the so called public key, while the other one is deposited in a safe place (private key). With the help of the so called RSA algorhythm and any key we can code a text but that coded text can only be decoded with the help of the other key.

In order to produce an electronic signature there is another technical tool needed and that is the hash-producer. This means that to any text we can unambiguously order a series of numbers comprising 32 hexa. This is the hash value[7] of the text. If we code this hash value with our private key then we have signed the document. The following formula expresses the production of the signature:

e-sign = Priv (Hash (text))

where the functions denote the private key encryption and the hash-production.

The recipient of the message has to do the following:

1. decode the hash-code with the help of the sender's public key,
2. produce the hash value of the message with the help of the hash coding algorhythm known to everybody,
3. compare his own hash value with the decoded hash-value.

If the two values are identical then the message was definitely sent by the sender and the transferred message is identical with the original one, therefore the signature is not false.

[5] The abbreviation contains the first letters of the three inventors' names.

[6] We can print it as a number onto our business card or 'hang' it up onto an Internet key-holding server.

[7] In principle 'only' 2^{128} different texts van have their own hash values. But this is a very large number, which we could not use up even if we were to write all our lives.

The sender can code the content of the message with the recipient's public key. This way the message can only be decoded by the recipient. And thus way we have solved the problem of encryption mentioned in the title of this part.

According to the Hungarian signature law[8] there are several types of electronic signature:

- simple
- with increased safety
- certified.

The only criterion applying to the first one is that the signing party has to be safely identifiable by it. The second type of signature basically corresponds to the one discussed above in more detail. In addition to its above description the law prescribes that the signature is made with a tool that can only be influenced by the signing party. A certified signature is an even more secure signature, the authenticity of which is guaranteed by a certificate.

The place where the certificate is made out has a key role in this process. This place is an authority, which is accepted by all clients and which - as a reliable third party - provides certificates based on requirements. In these certificates the characteristics of the person or organisation (e.g. name, unique data) are unambiguously connected to the electronic certificate containing electronic data prepared especially for them. The authority keeps the strictly defined rules for giving out a certificate and at the same time takes care of the records precisely, carefully and securely. (In this case security is meant not only in a technical sense but in the most comprehensive way, even including mechanical safety.)

This authority providing certificates is a place, which is audited at least once a year by a state-body predetermined by the law.

The so-called time-stamp is very important in the case of e-signature. This is so because in the case of signed documents the date is usually decisive. It is obvious that an electronic document containing a valid electronic signature can only be supplied with a date valid at the time of the signature. This valid date is supplied by the time-stamp service[9].

[8] Parliament accepted the 2001/XXXV. Act regarding electronic signature in June 2001.

[9] The problem of the time-stamp is even more complicated because one point in time can only be securely marked by the simultaneous use of two time-stamps.

Electronic payment tools

In e-commerce the value of sold products can most easily paid by electronic payment tools. Electronic payment tools however are not only used in e-commerce. The POS[10] terminals or the ATM[11] machines can also be regarded as such.

The real e-payments adjusting to the use of e-commerce are grouped in a recommendation by one of the respective EU committees as:

- remote access payment tools and
- electronic payment tools.

In the first case the user can access his/her own bank account and do different types of financial transaction on it. The electronic payment tool has to be rechargeable and capable of making payments.

In both cases the legal regulations are far-reaching. They regulate what should happen in the case of damages caused by:

- unaccomplished actions,
- demands not approved by the card-holder,
- faulty operation of machines,
- not adequate checking,
- selling or failed sales.

From a technical point of view e-payment has more than one solution. The simplest one is when the payment is made upon delivery, to the delivering personnel of the store in a traditional way, by cash. Instead of a delivering personnel the post can also be involved in the process. In that case the customer can pay by a postal voucher. Naturally, the same solution can be implemented by any carrier service not only the post, where the carrier company pays the electronic store afterwards.

It is also a expending method also in Hungary, that the consumer pays upon delivery with a bank card through the help of a mobile card reading terminal. This mobile terminal is connected via a mobile-phone to the bank's terminal, with which the Internet company has made a contract.

---

[10] POS (Point of Sales) – A place accepting cards. A POS terminal is a place, where one can shop and pay by an electronic card, since the system of electronic acceptance is developed. In the event of shopping there is an immediate connection between the computers of the business and that of the card providing bank. Therefore if there is not enough money on the bank account or if the banking card is cancelled shopping cannot take place.

[11] ATM (Automatic Teller Machine) – Bank-automate. The well-known machine is used primarily to receive cash without having to go to the bank. At certain machines one can also deposit money.

The most advanced method is when as a last step in the shopping the customer can pay through the computer. This happens by:

- a bank card or
- from a bank account.

Payment with a bankcard is based on the principle that the web site, where the customer initiates the payment, simulates[12] the operation of a POS terminal.

There are different types of bankcards:

- classical debit card, for example the old-style cards (e.g. Visa Gold, Eurocard/Mastercard Standard),
- electronic debit cards, like the newer types of debit cards (e.g. Visa Classic, Visa Electron),
- credit cards (e.g. Visa Classic Credit,
- Internet cards (e.g. Visa Internet)

Cards used for bank payment are usually credit cards or debit cards. The difference between the two of them lies with the bank account types that the bank opens for the client. Credit cards are issued only upon certain guarantees as this card enables the customer to withdraw money from an ATM even when the customer does not have the appropriate coverage on the account. This is possible because in this case money is actually withdrawn not directly from the customer's account. Most foreign ATM machines accept only credit cards as technically this is a simpler solution. When credit cards are used in a traditional way abroad then the customer usually has to sign as a guarantee for the charge. In the case of debit cards[13] (abroad) usually the PIN[14] code is enough.

Both card can be used for bank payment, but there are certain requirements that call for the use of cards made by even more modern technology, that is:

- electronically and
- with the use of raised print.

These cards contains a so called CVV[15] code which increases security. Also, on the raised print there is usually an identification number which is often asked for on the web-page.

---

[12] When a program simulates a machine then we use the term 'emulate' that is electronically simulate.

[13] This card is issued for checking accounts.

[14] Personal Identification Number.

[15] Cardholder Verification Value. This can only be found on newer cards where the card-issuing bank can identify the card user.

The traditional POS terminals call up the local telephone center through a simple phone-line and connect to the bank. The POS emulators, that is the POS terminals simulated by the computer, connect to the Internet and through that they reach the bank. Through a not secure web connection all the information on the card can be accessed by an unauthorised person. Therefore a supplementary security system needs to be used between the POS emulator and the bank in order to make the connection through the Internet secure and inaccessible to third parties. The technology for this is called SSL[16]. SSL is not only used in the case of payment but also when we send non-public data through the web from the client's browser to the server. (For example when we send personal data through a commercial site.)

SSL is the abbreviation for Secure Socket Layer, which encrypts the communication channel between the customers' browser (Internet Explorer, Netscape Navigator) and the remote server. Through the 128 bit encryption of the 'channel' the bank card data is inaccessible to others. According to the SSL standard for each and every banking transaction there is a different encrypting parameter chosen, which can be produced based on a random number.

An even greater security can be achieved in payment transactions based on the SET[17]-protocol, which was specifically developed for e-commerce. SET is a credit card payment method supported by Visa, MasterCard, Microsoft and Netscape. Here, in the course of the payment communication repeatedly encrypted sets of data travel (who buys, what and for how much) among the computers of the customer, business and the bank in a way that the information made available to every participant is only the part concerning him/her. The customer knows all three parts, but the business does not receive the customer credit card number. The data are encrypted by RSA-coding, that is public-key coding, therefore only the authorised person in possession of the private key can have access to it.

The SET developing society has established the SET Mark initiative. The experts of SET check the web sites on request and if they correspond to the SET protocol then they supply it with their trademark. This can enhance the trust of customers.

[16] There are also other technologies in use, but at the moment SSL is the most popular.

[17] SET – Secure Electronic Transaction

General contract conditions

In e-commerce there is tendency for the increasing use of the so-called 'form' contracts. This is the consequence of the fact that usually on the web there are uniform electronic forms.

These form contracts can make the process of contracting very fast, but at the same time do not provide for individualising the contracts. It should be noted, however that mostly there is no need to individualise and sometimes even no way. Still, the basic problem of this method is that:

- the party determining the contract has a stronger position than the other party moreover
- the party signing contract is not or hardly aware of the conditions in the contract.

Therefore the Civil Code – in accordance with the EU regulations – has been modified, taking the interest of the customer more into consideration. According to that these type of form contracts are only valid if the signing party had the possibility to learn the content of the contract and s/he explicitly and in a documented way accepted that.

The mouse-clicking methods of web e-commerce belong to the category of the so-called contract acceptance by indicative conduct. According to this new modification one click does not automatically result in the special conditions mentioned in an earlier phase becoming part of the contract. Therefore there is an urgent need to codify the written documentation of the electronically made contracts.

6.5.7. The infrastructure of e-commerce

The fundamental reason and cause for the introduction of e-commerce lies with a solution for the exchange of information both within and organisation and between organisations that can go beyond time and space. Internet and within that the web technology provides the concrete method for this.

The general model of the system

E-commerce makes use of a rather diverse technical system of tools, where besides successful, older systems such as e-mail or EDI[18] (Electronic Data Interchange) new and recently developed systems are also being applied. The model and the

[18] EDI is a so-called electronic mail application, where in determined internal format e-commercial transactions (e.g. ordering, bank transfer) are implemented between the partners in the form of e-mail.

architecture of e-commerce is based on traditional forms, where the aim is to reflect the transactions between and together with the different actors, who are the:

- customer / client
- deliverer
- product
- service etc.

In order to handle the complicated relationships it is useful to introduce the general ‘broker’. The concept of the electronic broker captures a service, which provides the most attractive product for the client at a given time and place. The strength of applying a broker system is that it guarantees the widening and spreading of the market for the customer by informing the customer only about the advantages of the market while at the same time s/he does nor have to indulge into its complexity and complicated nature. Broker systems are also advantageous for the delivering company as they receive the needs of the customer in a simple way without having to deal with client services in depth.

According to that the broker is in touch with three actors, with:

- the client,
- the deliverer and
- other brokers.

In the famous COBRA (Common Object Request Broker Architecture) project there was an attempt to divide the activities of business actors into phases in a way that the commercial activity could be examined on an abstract level. On that basis in the common model difference is being made among:

- encounters,
- transactions and
- follow-up work.

Through the encounters the appropriate information reaches the interested party and a successful encounter brings about a transaction. Encounters are connected to a number of prerequisites and its final stage means payment, delivery and other exchange of information. The phase after the transaction deals with possible complaints or questions of responsibility.

The electronic broker architecture can only be realised with the help of the appropriate computer science methods. The difficulty of the solution arises from the heterogeneous contents and forms, the insufficient and not unified meta-descriptions, the different technologies and many other issues.

Computer science solution

The e-commerce application systems use the Internet and naturally contain web systems as well. In the historical development of the web three levels of web-use can be differentiated:

- simple,
- interactive and
- object oriented.

The e-solutions are almost exclusively based on object oriented web. The object webs are web systems, where the web server – unlike the earlier web solutions – is object oriented. This category includes the servlet based system, the JSP (Java Server Pages) developed by SUN or the ASP (Active Server Pages) by Microsoft.

Servlet is the object oriented version of the web CGI[19], which however has radically given up on the rather simple technique of CGI-solutions. Servlet is in fact a Java class, where every time a new request arrives from a customer a new thread is being created, which then processes the request. The high level ability of Java[20] to handle exceptions assures the unambiguous and determined processing of all expected and unexpected system activity, which at the same time excludes the possibility of break-ins resulting from the use of not checked CGI scripts. There are further advantages to the use of Java-based servlet machines. These system can be easily scaled, as a consequence of their OO nature they can be relatively quickly developed since on the basis of the ready object the creation of new ones does not cause a big problem.

The solutions for the client-pages of the web are also important elements of e-application. Today the following systems count as pioneering technology:

- so-called dynamic HTML and
- client page Java.

The dynamic nature of dynamic HTML has largely been increased by the introduction of style-pages. With the help of CSS language to every single element of HTML we can assign unique styles, moreover these styles can be altered both in time and

[19] CGI – Common Gateway Interface. This is a system, which helps to evaluate and process the data submitted on a web page. Technically there is a program started on the web server and it results in a dynamic web page, which then goes back to the user.

[20] Java is the most popular program language nowadays. Many consider it the XXI. century version of C. It belongs to the so-called object oriented program language family. ‘Operations’ (like adding or searching) as such do not exist in general. They can only be understood as part of an object or object family (e.g. whole numbers, 27-character-long character chain).

place. The changes in the web document can also be followed by the DOM model. The different activities of the client (such as the clicking of the mouse or the moving of the mouse to a certain pixel position) can be evaluated as events realised and which can be easily processed by JavaScript or Jscript usually. The processing can result in a change of appearance on the web site or other alterations.

In the traditional structure of HTML special elements can be built in on the JSP and ASP pages. In the case of JSP these elements are in fact Java language program parts, differing declarations or object calls. With the help of this the web page is capable of realising more complicated and complex processes. It should be added that JSP becomes a servlet on the server, therefore this technology is a form of server solutions.

On the client's side, that is in the browser, Java systems can be used differently. These are called applets, that is smaller programs. Applet is in fact a Java program with a fixed format, which runs on the Java machine built into the browser. With its help a user-friendly, intelligent environment can be formed. Security is further increased by the fact that the developer of the applet can sign it the same way as an electronic document would be signed. This way it can be determined who the maker of the applet is and if we trust that person, the running of the applet can be permitted.

The so-called component-principle system formation is the computer science mapping of the CORBA principle. According to that a big system – most e-commercial applications can be regarded as such – consists of component and these components have a well-defined interface. These interfaces are used from a distance, that is through the net as remote processing invocations. This solution fits the object oriented view very well, where the objects – be those a remote ones – expect messages, to which they send back their replies depending on their internal state. Therefore - and for other reasons as well – the components are programmed in OO language and thus Java has an exceptional significance in the developing.

Today there are two standards for processing invocations. One is actually Java RMI, which is the so-called remote method invocation[21]. The CORBA-IIOP standard is also often used. The advantage of using the CORBA-IDL style OO system is that it does not require a pure Java system, it can also function in a heterogeneous environment. Today there are several such systems developed, which are although not Java but it still has a Java-RMI interface and therefore the connection can be solved.

[21] In Java the objects can be called not through procedures but through the so-called methods.

XML (eXtendable Markup Language), which is becoming an industrial standard, is the most suitable tool for the description of the format of the exchange of information between the components. This language evolved from the HTML of the web, but it is much more flexible since the actual structures are invented and defined by the user. For the processing of XML structures freely usable system components were developed, among which the Java based SAX (Simple API for Xml) is the most well known.

In e-commerce applications the handling of the components in adequate sequence – which is called business logic- should practically be solved in separate units. Following the server-client co-operation scheme, a so-called application server has been developed as a framework. The application servers contain a Java virtual machine and therefore they are ready to run Java setvlets or other Java moduls. These modules then call the individual components. The communication is realised through the already mentioned Java-RMI based remote system invocations. With the help of these system invocations the back-up databases and other service – e.g. address books, document handling systems - are accessed and used.

References

[6.5.1] Antal & others: JAVA 2 - Guide for programmers

ELTE TTK H.A., 2000

[6.5.2] Garfinkel: Web Security & Commerce

O'Reilly, 1998

[6.5.3] Mojzes-Talyigás: Elektronic commerce

Typo Art Studio & Kelisprint, 2000

[6.5.4] Benchmarking Telework and E-Commerce in Europe
ECATT Final Report

http://www.ecatt.com/ecatt/

IST Programme

New Methods of Work and Electronic Commerce

http://europa.eu.int/ISPO/topics/i_ecom.html

6.6. Advanced teleservices

Gábor Magyar, author

István Cselényi, reviewer

Technologies create possibilities whose evolution, when realized, is then drive by a combination of economic and social factors. *Information and Communications Technologies* (ICTs) are already an integral part of our daily life, providing us with useful tools and services in our homes, at our workplaces, everywhere. The Information Society is not a society far away in the future, but a reality in daily life. Based on existing technologies in telecommunications, informatics and media next generation of user-friendly, cost-effective and interoperable general interest services are emerging, meeting user demands for flexible access, for everybody, from anywhere, at any time.[22] Application areas cover practically the entirety of our everyday life. e-commerce, teleworking, e-government, content industry and telehouses are explained in separate chapters. Here You can find brief overview of learning/teaching, healthcare, cultural and transport applications – focusing on technological constituent rather then socioeconomic and ethical aspects.

6.6.1. Learning/teaching applications: eLearning

The digital age has spawnes an overhelming mass of raw information that is frequently difficult to retrieve and to use. Network technology has enabled a proliferation of customized and timely educational tool that optimizes investment in human capital: **eLearning solutions facilitate the delivery of the right information and skills to the right people at the right time.** (What is more sophisticated, than the „anyone, anyplace, anytime" model!)

eLearning will be critical to the success of individuals, organizations, communities and economies in the dawning knowledge economy. The globalization of the economy, shortage of skilled workers, free agent mentality, new flexible work situations (e.g. telecommuting) have created problems not easily solved by traditional education. Companies in this space are addressing these problems using technology-based learning

[22] A series of European examples can be found in the EU's IST Programme, ("Systems and Services for the Citizen")

resources targeting the academinc, corporate and consumer spheres with managed, interactive, user-centric education tools.

What is the role of technology in future education?

Web-based collaboration tools enable progress in far reaches of the world to share an educational experience, while sophisticated learning management systems allow a user's progress and preferred instructional style to be tracked so that adjustments can be made to a personalized curriculum in order to effectively address learning objectives. Advancements in the overall capabilities of networked technology (broadband access, wireless access) will serve as a catalyst for acceptance of eLearning technologies on a more widespread basis in all markets. Such improvements will make user concerns about capabilities of the technological medium secondary to actually providing meaningful educational experiences.

Typical applications are:

- Tele-education for vocational training: small businesses, training centres of large companies and public authorities.
- University-level eLearning offers the possibility of the use of interactive teaching material and remote dialogue between students and lecturers.
- Networking between primary schools and secondary schools will help to enrich curriculum content and provide access to new, more sophisticated teaching aids.
- eLearning for job-seekers will give the unemployed the chance to upgrade their professional skills or retrain individually at home or in groups.
- eLearning provides opportunities for the general public, particularly the elderly, the disabled, the rural population too. They will gain access to personalised education in their own homes and will be able to communicate with other students.

The emergence of a set of open industry-standards will be a critical turning point for the industry as a whole. *eLearning standards* will allow learning content to be easily accessed and reused in various formats and will enable the interoperability of learning technologies from different vendors.

Content Development

The key planning activity in the design phase is the creation and articulation of specifications for content development. Specifications need to cover the range of development concerns, including technologies to be used, document templates, markup definitions (for example, the Document Type Definition in SGML-based software), editorial standards, modularity requirements, structural rules, and the level of granularity desired. If

these components are not decided in advance, developers may waste time re-creating content to meet the needs of either the project at hand or reusability requirements.

Once specifications are established, they can be reused or tailored for individual projects.

The key for deploying learning objects effectively is to provide ways for the learner to contextualize the information. Without context, learning objects can be confusing, misleading, or utterly meaningless. Context is the second path for personalization of objects (after adaptive selection of appropriate objects based on individual needs). Providing the original context of an object will often be inappropriate (and in many cases defeats the adaptive purposes of breaking instructional material down into smaller objects). Yet how much context is enough? Perhaps a better question would be: How can context be scalable in expanse and type, so that the learner can decide how much is needed?

Constructivist theories and active learning theories have helped educators understand the way learners actively create meaning by exploring, experimenting, testing, and applying knowledge in self-directed and collaborative fashions (rather than in a predetermined course of study). Use of learning objects will empower online learners in unprecedented ways by enabling them to participate more actively in the contextualization of information. In this paradigm, context is not something that is simply provided to a learner. Instead, contextual information has two functions: to orient objects to their original or most likely contexts, and to provide cues for learners to apply their own meanings and contexts to the information.

Whatever development environment and tools are used, sound instructional design will remain important both for customized development and for template-based development. The combination of thoughtful planning with intelligent deployment of advanced authoring tools will result in myriad benefits for both content producers and learners. The most successful learning object delivery systems will be able to provide not only learning object content, but relevant and meaningful context, as well.

Learning objects

Designers and developers of electronic learning today are being presented with a new content development landscape. Learning technology standards organizations are moving towards open and industry-wide standards for *learning objects*. As learning content developers look at these initiatives that focus on packaging, identifying, and exchanging content, they are bound to ask, "What does this mean for me? How will my work be different in the future?"

Most electronic learning content is currently developed for a specific purpose such as a course or a situational performance intervention, and not for the sake of populating an objectbase (a collection of learning objects, typically contained or referenced in a *database*). Why would designers wish to add a layer of complexity to their work by including object capability in their design? The reason is that this effort - in most cases - will pay off many times over (in terms of costs, development time, and learning effectiveness). The object approach can satisfy both immediate learning needs - such as a knowledge-based or skills-based course - and current and future learning needs that are not course-based.

There are two requisite components of a learning object: the *object content* and its *metadata* tag. Descriptions and keywords provide some degree of context, yet ideally there are additional contextualizing options. As software developers race to produce authoring and tagging tools, it remains to be seen what various context-enriching options will be available. The most desirable tools will permit scalable contextualization so that learners can control the extent to which context is presented with content. A key issue is *reusability*. In an environment in which context is scalable and adaptive, the ideal Reusable Learning Object (RLO) content is

- modular, and transportable among applications and environments
- nonsequential
- able to satisfy a single learning objective
- accessible to broad audiences
- coherent within a predetermined schema (so that a limited number of metatags can capture the main idea or essence of the content)
- it can be repurposed within a different visual schema without losing the essential meaning of the text, data, or images.

When learning content is created both for immediate purposes and for use as RLOs, designers and developers must enlist a sort of "double vision." This entails conceptualizing content as part of a larger whole (such as a course) and as stand-alone information at the desired level of granularity. These do not have to be conflicting activities, though to accomplish both successfully and efficiently requires thoughtful planning. RLO content needs to be grounded in solid instructional design, so the new landscape of learning objects will welcome the efforts of experienced instructional designers.

The main arguments for designing and developing material to be reused as learning objects are as follows:

Flexibility. If material is designed to be used in multiple contexts, it can be reused much more easily than material that has to be rewritten for each new context. It's much harder to uncouple an object from the context of its parent course and then recontextualize it than it is to contextualize as part of design and development.

Easy content management. Metadata tags facilitate rapid updating, searching, and management of content by filtering and selecting only the relevant content for a given purpose.

Customization. When individual or organizational needs require customization of content, the learning object approach facilitates a just-in-time approach to customization. Modular learning objects maximize the potential of software that personalizes content by permitting the delivery and recombination of material at the level of granularity desired.

Interoperability. The object approach allows organizations to set specifications regarding the design, development, and presentation of learning objects based on organizational needs, while retaining interoperability with other learning systems and contexts.

Competency-based learning. Competency-based approaches to learning focus on the intersection of skills, knowledge, and attitudes within the box of core competency models rather than the course model. While this approach has gained a great deal of interest among employers and educators, a perennial challenge in implementing competency-based learning is the lack of appropriate content that is sufficiently modular to be truly adaptive. The tagging of granular learning objects allow for an adaptive competency-based approach by matching object metadata with individual competency gaps.

Value of content. From a business standpoint, the value of content is increased every time it is reused. This is reflected not only in the costs saved by avoiding new design and development time, but also in the possibility of selling content objects or providing them to partners in more than one context.

Constructivist Teaching
Constructive pedagogy (as a counterpart to behavioristic pedagogy) stresses the importance of teaching (which aims at the generation of understanding) versus pure training for performance (often geared at perfectly solving textbook problems). Knowing as an adaptive activity constructs a set of successful/viable concepts, models and theories relative to a context of goals and purposes. Learning requires self-regulation and the building of conceptual structures through reflection and abstraction. Problems are not solved by the retrieval of rote-learned "right" answers.
The KBS Hyperbook is a system which uses explicit conceptual models and metadata to structure and connect external data. When these external data are pages on the WWW, the corresponding conceptual model takes the role of an information index and determines the navigational structure between these pages (corresponding to one or more views on the external data). The conceptual model also serves as a schema for the integration of new pages (similar to the role of a database schema).
Central to the structural model is the concept of a *semantic information unit* (SIU) whose instantiations contain the main information units contained in the hyperbook. The set of SIUs is used for modeling the application domain. Relationships between SIUs are modeled by several semantic relations, which structure the knowledge referenced by the SIUs, for example to relate general knowledge to specializations, related concepts, etc. Semantic structures that emerge for domain modeling are e.g. taxonomies based on inheritance hierarchies and more general domain ontologies including arbitrary relations.

6.6.2. Healthcare applications

New generation healthcare applications cover new generation computerised clinical systems, advanced telemedicine services and health network applications to support health professionals, continuity of care and health-service management, and intelligent systems allowing citizens to assume greater participation and responsibility for their own health.

Professional healthcare systems enhance the ability of healthcare professionals, for prevention, diagnosis, care and rehabilitation, such as intelligent systems for non-invasive diagnosis and therapy, intelligent medical assistants, advanced medical imaging, and advanced telemedicine applications. *Virtual healthcare facilities* offer single-point-of-entry services, using high-speed secure networks and applications for linking emergency services, hospitals, laboratories, pharmacies, primary care and social centres and the home for continuity of care. *Health service workflow management and re-engineering* creates new generation electronic health records and cards for sophisticated health-data objects, personal health systems. Affordable and user-friendly systems can be disseminated for personal health monitoring and fixed or portable prevention systems, including advanced and affordable sensors, transducers and micro-systems.

In general: tele-systems and applications support healtcare in all contexts. ICTs and related technologies are expected to yield a host of benefits for the development of health services, including the collection and analysis of information, the identification of high-risk groups, health services to remote and under-served groups, support for citizens' own health-promoting activities, etc. Due concern needs to be given to the protection of confidential health data in ICT-based systems, and to the requirement of reviewing the ethical codes of health professionals in the light of ICT-based health practices. Moreover, user-friendly and certified information systems are requiered for supporting health education and health awareness for citizens as well.

What is the role of technology in future healthcare?

Telemedicine and the new remote medical services are based on the technology for the storage and mobile communication of digitalised medical information. Telematic links, coupled with the use of microprocessor cards (smartcards) serve to access the networks and transfer the essential elements of medical records, connect patients, general practitioners, specialists, laboratories and/or hospitals.

Typical applications are:

Computerisation, exchange of and shared access to medical records. Clinical and administrative data about patients can now be accessed and shared in real time by authorised parties, such as hospitals, insurance companies and mutual insurance societies, public health and social security institutions, laboratories, practitioners and/or healthcare workers. Administrative and financial procedures are simpler, quicker, safer and cheaper as a result.

Telematics networks based on common communication standards can ensure the interoperability of multimedia workstations, diagnostic aids, on-line consultation of medical databases and records/archiving systems, i.e. national health systems on an international scale.

The introduction of mobile telematics services for first aid.

The development of hospital information and communication systems linking all departments, such as general medicine, surgery, accident and emergency, intensive care, radiology, pathology, etc.

Medical imaging: the transfer and/or remote visualisation of medical images such as X-rays, scans and electrocardiograms.

Remote consultations: using videophones specialists working at a distance can carry out initial examinations, produce a diagnosis and help doctors in remote regions.

Systems for administering medical prescriptions make possible remote analysis of the patient's personal records in conjunction with the specific details of prescriptions.

Routine examination and monitoring at home of patients with restricted mobility, such as pregnant women, newborn babies, the elderly and disabled.

Interactive monitoring of surgical operations. Access to telematics networks for healthcare professionals is essential in the supply of first aid at the scene of accidents or natural disasters in remote regions.

The establishment of networks of transplant organ and bone marrow banks increases the chances of finding suitable donors (European EMDIS Project).

Broadband videoconference and telehealth applicationsA broadband videoconference facility was developed in the frame of a EU NICE project to create interactive collaboration platforms, to connect large number of endpoints, via heterogeneous network (IP over ATM, ISDN, Internet, satellite). TCP-UDP/IP protocols were used, for more than two participants UDP multicast were recommended. Bandwith ranged between 256 – 6000 kbps.

Interactive Site (IS) is a site that could interact (send and receive audio/video/data) between themselves using a special software. Watch Points (WP) is a site that could only receive audio/video/data sent from the Interactive Sites. Network node (NN) is a site which aggregate traffic from the leaves towards the root and perform traffic broadcast from the root towards the leaves. A Control Site (CS) is required to remotely control the configuration of all the applications located at the Interactive Sites.

Cardiovascular Health Care Telelink to Hungary Based on the previous experiences a special telehealth network were set up between the University of Ottawa Heart Institute, the Institute of Cardiology in Budapest and the Budapest University of Technology and Economics, Departement of Telecommunications and Telematics. The system included a telehealth platform that collects patient data, a signal processor that encodes the audio, video, and data signals and then sends it through the network to another telehealth platform.

The telehealth services were as follows- access remote sites with single mouse clicks
- conduct interactive case consultations
- transmit and store medical images like x-rays, MRIs and echocardiograms
- capture images from a video camera, medical instruments or high-resolution film scanner
- multimedia consultations
- perform real-time ultrasound
- control both local and remote camera
- transmit heart and lung sounds

Disabled and elderly

In various cities and regions across Europe, projects have been set up which use ICTs in innovative ways to help older and disabled people. The importance of ICTs as a tool to help promote integration of older and disabled people has also been recognised in a number of EU programmes. HANDYNET, for example, is a European-wide computerised information and documentation system on technical aids for disabled people, and research and development into the use of ICT products and applications to help disabled and older

people in their daily lives is currently being promoted under the EU's TIDE Programme (Telematics for the Integration of Disabled and Elderly).

6.6.3. Cultural applications: access to knowledge

One of the main challenges facing the information society is to give as many people as possible access to the cultural and intellectual heritage available in memory institutions (public libraries, archives, museums). This may mean upgrading services in public libraries or creating new interactive systems for using archive collections, but in either case digital technologies for the multimedia processing and dissemination of information are being used to open up access for a much wider public to our cultural and intellectual resources.

Typical applications are:

computerised library management: computerisation of documentary sources,

information storage and accessibility: electronic archiving, electronic distribution of documents, development of interactive tools to enable users to consult collections, virtual museums

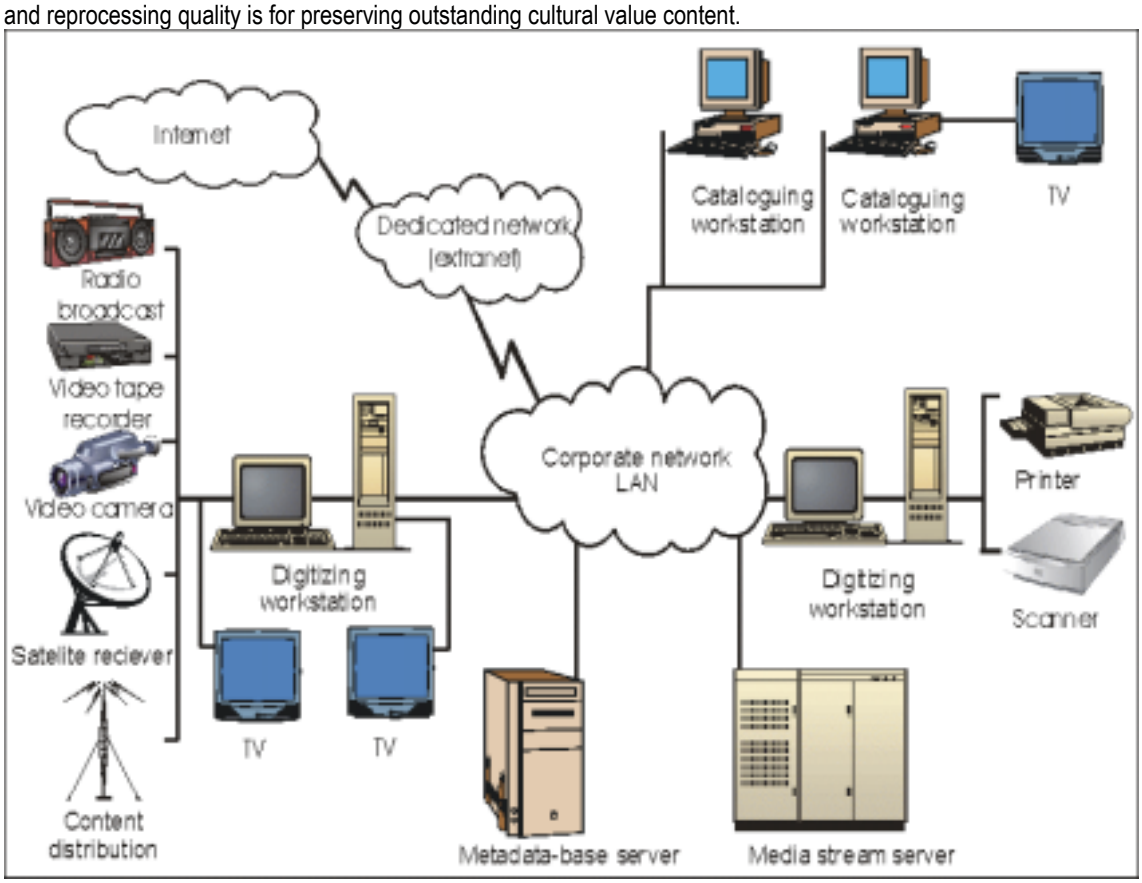
value-added library services: establishing library networks, connecting libraries and publishers, integration of public library services and other multimedia services for the general public, such as eLearning.

Searching/browsing systems: intelligent searching systems for information retrieval in different memory institutions (using standard meta-database schemes)

National Audiovisual Archive

The National Audiovisual Archive (NAVA) pilot project aimed at creating the first national legal deposit audiovisual broadcast archive in a digital environment in Hungary. The goal was to create a deposit of the national audiovisual heritage that is compatible with international technological and metadata standards so that it can be eventually connected to any international open archival system. The main technological goal has been to create a full on-line digital environment, which includes a high speed connection between the broadcasters (television and radio stations) and the archive on the one hand, and a dedicated on-line network connecting the archive and big libraries on the other hand.

The core of the archive is a large video server that functions as a real-time digitizer of the incoming audiovisual material. After processing the metadata the audiovisual content and the metadata is stored separately. The audiovisual content is stored in three different qualities: the entire material is stored in *browsing quality,* and *broadcast quality*, a small portion of it is stored in *reprocessing quality*. Browsing quality is used for dedicated consultation purposes, broadcast quality is for preservation purposes and reprocessing quality is for preserving outstanding cultural value content.



The metadata structure of the NAVA is based on the Dublin Core scheme so the core of it applicable in all kind of archival activities. On the one hand this means that NAVA database is connected and is searchable for all sort of documents of cultural heritage, on the other hand this should make possible to standardize different audiovisual databases. This way NAVA model functions as the data exchange center for audiovisual archives and the connection center for all Hungarian cultural heritage databases.
(NAVA pilot project: Budapest University of Technology and Economics, URL: www.nava.hu)

6.6.4. Transport

ICT offers instant solutions to some of the public's traffic problems. The various telematic services make it possible to use the road network more effectively, to improve road safety, increase the efficiency of public transport and the haulage industry and reduce environmental nuisance and pollution. In addition, these integrated networks can be used to develop new value-added services (e.g. based on computerised tourist information), which will help to create new jobs in the very short term.

What is the role of technology in future transport?

ICT in transport can serve various objectives: to increase safety, capacity or convenience within individual modes of transport, to shift traffic from one mode to another, more environmentally friendly one, or to avoid traffic altogether. However, only local short-term impacts have partly been assessed but the longer-term impacts, in particular with respect to changes in the traffic patterns and modal split, are difficult to assess. Only some of the ICTs are expected to have an impact on traffic levels at all. In the long term, relief from traffic problems through transport telematics can only be temporary until the efficiency and capacity gains are offset by the general trend of traffic growth

While many transport telematics technologies have proven their technological feasibility, a number of non-technical challenges remain on the way to exploiting their potential. Firstly, there is the issue of standardisation. In several areas largely incompatible technologies are likely to compete for market penetration. Examples are the different communication technologies for traffic messages and dynamic route guidance in road transport, or the tracking & tracing technologies in intermodal freight transport. Another problem is the slow adoption of available ICTs in a number of application areas. The potential long-term benefits may be high, but it is locked by the inertia of current procedural and organisational structures.

Typical applications are:

Automatic navigation systems for cars. Such systems are based on the automatic exchange of information in real time between the vehicle and a "nerve" centre supplying the data on the one hand and between the vehicle and the driver on the other. Information provided may be about the traffic flow, the state of the roads, accidents or weather conditions. The information is picked up via a network of short and medium-range communications beacons, electromagnetic sensors and cameras and relayed by video or telephone cables to a traffic centre which, in turn, relays it to motorists by radio, digital display on the car radio or on-board computer. On-board information projects can be applied to urban traffic, national and cross-border trunk roads, the interconnection of traffic management centres and the management of public transport systems.

Freight management. EDI and mobile communications can be used to improve the management of vehicle movements, the trans-shipment and monitoring of goods using combined road-rail transport, the transport of dangerous loads and customs inspections.

Electronic payment. Electronic toll cards, electronic payment of urban parking charges and electronic ticketing systems are ways of reducing investment and operating costs for operators, cutting theft and vandalism and promise greater flexibility and convenience for users.

Travel information is now available in many cities before departure, at home or in the office, at bus stops, metro and train stations, or via fixed or movable signs along the route. It has become a factor in attracting people to public transport and prompting drivers to leave their cars in carparks on the edge of town.

Challenges ahead

Private car: The most visible telematics elements in road transport are variable message signs. On motorways they are used to adjust speed limits and traffic flow to the traffic conditions or to warn of incidents. In urban areas they are often used for parking management and park & ride schemes. A reduction in traffic accidents and road capacity increases could be proven in many cases. Static route guidance systems, based on digitised road maps and combined with a satellite or other positioning system, have been available as options for luxury cars for years. Real-time traffic information which is necessary for dynamic route guidance can be conveyed to cars either by Radio Data System/Traffic Message Channel (RDS/TMC), Digital Audio Broadcasting (DAB), mobile communication (GSM) or short range communication through beacons installed along the road infrastructure.

Public transport: Static railway timetable information is now integrated and available for the whole EU for inter-city, and in some countries also for regional trains.

The Netherlands have already managed to realise a system that also integrates the timetable information of all public transport services, be it train, bus or tram.

The problem is not the increased data volume of such a system, but the degree of automation and quality that can be achieved when integrating timetable information from numerous national, regional and local public transport providers. Names of rail and bus stations need to be harmonised and the interconnections between two different modes (e.g. inter-city train and bus) at the same station need to be defined (e.g. 'five minutes walk; elevator available for the handicapped and elderly'). This harmonisation is a first precondition for being able to propose optimal door-to-door transport chains.

For fully intermodal door-to-door information system the challenges go even further:

- integration of pricing, booking and ticketing for the whole journey;
- route optimisation not only based on the shortest travel time but also on the route with the fewest interchanges or the lowest total fare;
- integration of road directories in order to match stations/bus stops to destinations or frequently visited points of interest;
- incorporation of real-time information about public transport service interruptions, delays or deviations in order to move from a static to a dynamic information system.

Concerning operational aspects of public transport, localisation technologies and two-way radio and data communication systems are now being widely introduced for the management of fleets of public transport vehicles and to feed real-time customer information systems (e.g. indicating expected arrival times).

Railways: A major EU-wide telematics project in the railway sector is the European Rail Traffic Management System (ERTMS). Its prime objective is the harmonisation and advancement of procedures and technologies for train control and management across the heterogeneous national railway networks in order to achieve a higher level of interoperability. In its final and third stage, the ERTMS should allow for a 'moving block' based operation through enabling the self-location of the train and signal-less operation, via ground-to-train continuous or semi-continuous radio transmission. Such a type of operation has the potential to significantly increase the rail network capacity by reducing the necessary safety distances between trains. Satellite systems - which have already been used successfully in the USA for train control - are regarded as a very flexible solution, because they make any trackside control and communication infrastructure unnecessary.

Air traffic: Air safety is put at risk because of widespread dependence on short wave radio communication systems for the communication between air traffic control and aircraft. Messages get distorted or cut off leading to misunderstandings that can cause mid-air collisions of aircrafts. However, the technologies to overcome this problem are already available:

- Automatic Dependent Surveillance Europe (ADS-E) provides aircraft location for air space controllers while aircraft on trans-Atlantic flights are outside the radar range;
- the on-board Traffic Alert and Collision Avoidance System (TCAS) can inform the pilot of other aircraft within a range of 25 km and recommend evasive manoeuvres (the system is already obligatory in the USA);
- for the crowded airspace above airports, there are systems which improve short wave communication.

Port management and maritime transport: The development of port and maritime information systems goes in two directions: one direction is the creation of ICT platforms for electronic business networking. It connects heterogeneous computer systems of different parties of a port community, so that they do not need to homogenise their data processing formats and systems. The second direction is the standardisation and storage of information so that it can be used by different port communities.

Substitution of traffic by communication

Emerging information and communication technologies have given rise to the hope that these technologies will make at least part of the transport of persons or goods unnecessary, and that physical traffic can be substituted by electronic traffic. In reality it seems that the increased use of telecommunications will lead to effects on traffic in two opposite directions: reduction in and induction of traffic. (E.g. tele-workers not only save journeys but also the time used for travelling. There is much that indicates that at least part of the time 'gained' here will be spent on leisure activities, thus creating additional traffic.)

The use of telecommunications makes many activities less dependent on location and thus creates additional freedom inthe selection of location. This could reinforce both desired settlement developments (e.g. integrated or distributed use, economic development of rural and less favoured regions), but also problematic developments such as spatial-functional separation linked with a further steep increase in dispersed traffic flows.

For the substitution of physical traffic, it is possible to distinguish business and private applications. With regard to business applications, special attention should be given to tele-working. This has an influence on (daily) commuter traffic. The success of tele-working depends, however, less on the wishes of transport policy makers than on the interest of businesses and their employees. In addition, the employers achieve enormous gains from not having to erect and maintain buildings. Flexibility is increased as new work contents are hardly compatible with the rigid eight-hour working day. Employees can be recruited from a wider geographical area.

Besides tele-working at home, commuting to a tele-commuting centre closer to home is a strategy to reduce traffic. Tele-commuting centres could be satellite work centres of an enterprise close to the location of the employee's home, or they could be neighbourhood work centres, in which employees from a number of firms share a common infrastructure.

The emergence of IT-based multi-modal door-to-door transport service
Odessey exemplifies the emergence of new services through integration. It is an innovative multi-modal (i.e. using different means of transport) door-to-door transport service for (primarily) business travellers. Partners in Odessey are the car rental company Avis, Rent-a-driver (specialised in providing chauffeur services), the network local and regional taxi firms, (subsidiaries of the Dutch National Railways). Odessey offers transport brokerage services, both on the supply side (taxis) and on the demand side (travellers). The product boils down to integrated transport management services for individual business travellers. The service delivery system is based on:
- a chain management system arranging tailor-made trips from one postal code to another;
- a call centre for communication with customers - ideally through mobile phones - and the different transport services and for travel management - in case of delays the travel plan can be accommodated.
Each customer gets a reservation voucher, which also functions as a travel warrant in combination with the Odessey-card (a chipcard).

References

[6.4.1.] Building the European Information Society for us all. Final policy report of the high-level expert group. April 1997. European Commission, Directorate-General for employment, industrial relations and social affairs. Unit V/B/4

[6.4.2.] eEurope. An Information Society For All. Communication on a Commission Initiati e for the Special European Council of Lisbon, March 2000

[6.4.3.] Green paper on the convergence of the telecommunications, media and information technology sectors, and the implications for regulation. European Commission, Brussels, 3 December 1997

[6.4.4.] Henze,N., Naceur, K., Nejdl, W., Wolpers, M.: Adaptive Hyperbooks for Constructivist Teaching. URL: www.kbs.uni-hannover.de/hyperbook/

[6.4.5.] Learning in the information society. Action plan for a European education initiative. 1998.

[6.4.6.] Living and working In the Information society: People first. Green paper

[6.4.7.] Longmire, Warren: A Primer on Learning Objects. Learning Circuits, URL: www.learningcircuits.org

[6.4.8.] Magyar, G.:A telematika és a fenntartható társadalom. Magyar Tudomány, 1998/11, pp 1298-1310.

[6.4.9.] Magyar, G., et al.: Metadata System of National Audiovisual Archive in Hungary. Invited Paper. 20th Conference of the Audio Engineering Society: Archiving: Restoration and New Methods of Recording. Budapest, 5-7 October 2001.

[6.4.10.] Magyar, G.: Networked Society and the Sustainable Evolution. Invited Paper. International Conference on Infocommunication Trends. Budapest, 11-12 Oct. 2001. Proceedings: CD-ROM

[6.4.11.] Magyar G.: "Fenntartható" lesz-e az információs társadalom? INCO, 1999/1. http://www.inco.hu

[6.4.12.] Mediasite: Net-Based Services Supporting Integrated Applications in Virtual Community, Product Engineering and Net-Based Learning.
URL: www.itnorrbotten.se/projekt/mediasite/index.shtml

[6.4.13.] Nicola Henze, Kabil Naceur, Wolfgang Nejdl and Martin Wolpers: Adaptive Hyperbooks for Constructivist Teaching. in KI-Themenheft vol. 4, 1999: Intelligente Systeme und Teleteaching.

[6.4.14.] Nicola Henze and Wolfgang Nejdl: Bayesian Modeling for Adaptive Hypermedia Systems. In *ABIS99, 7. GI-Workshop Adaptivität und Benutzermodellierung in interaktiven Softwaresystemen*, Magdeburg, September 1999. http://www-mmt.inf.tu-dresden.de/joerding/abis99/EndPaper/henze_16/henze_16.html.

[6.4.15.] Nicola Henze, Wolfgang Nejdl, and Martin Wolpers: Modeling Constructivist Teaching Functionality and Structure in the KBS Hyperbook System,CSCL'99: Computer Supported Collaborative Learning, Standford, USA, Dec. 1999.

[6.4.16.] Peter Fröhlich, Nicola Henze, and Wolfgang Nejdl: Meta-modeling for hypermedia design. In *Proc. of Second IEEE Metadata Conference*, Maryland, September 1997

[6.4.17.] Ruttenbur, B.W., Spickler, G. C., Lurie, S.: eLearning – The Engine of the Knowledge Economy. Morgan Keagen, New York, 2000. 07.

[6.4.18] Singh, Harvi: Achieving Interoperability in e-Learning. Learning Circuits, URL: www.learningcircuits.org

[6.4.19] Study on the Use of advanced telecommunications services by Healthcare Establishments and possible implications for telecommunications regulatory policy of the European Union. Empirica, Gesellschaft für Kommunications- und Technologieforschung mbH and Work Research Centre. Bonn/Dublin October 2000.

[6.4.20.] The situation of Telecommunications Services in the Regions of the European Union. Gallup Europe. 2000. URL: europa.eu.int/ISPO/infosoc/telecompolicy/en/EOStudy/Resid/2br.htm

[6.4.21.] Wayne Hodgins, Marcia Conner: “Everything you ever wanted to know about learning standards but were afraid to ask.” LiNE Zine’s Fall 2000 issue.

6.7. Distance work

Judit Talyigás, author

László Z. Karvalics, reviewer

6.7.1. The concept and breakdown of distance work

The concept of distance work spread in the framework of the issues of information society during the last fifteen-twenty years. Of course, the concept is not new, but its contents are. In the past, workers identified themselves as distance workers who met their employer once a week or once a month, and worked at home on the allocated, usually simple assembly, sometimes sowing, and other tasks. Translators were also distance workers, who took home manuscripts written in foreign languages from publishers and translated them into nice Hungarian language at home. The flying reporters were also distance workers, working far away from the employer and editorial offices. And, perhaps they were the first to send the product of their work, their reports, to the employer using technical devices, telex, telephone, fax and later Internet. In other words, the result of the work "travelled" from the employee to the employer.

The contextual change in the concept of distance work used in traditional and information society is that in the first case, the employee is far away from the employer, but the employee brings and takes the product of the work in the majority of the cases. The basis of the new concept of distance work is that the subject of the work reaches the employee electronically, and the result of the work, the product is returned to the employer also electronically. Although the belief, that distance work means only work at home, is obviously wrong, we still need to note that mainly this is the case. Of course, there are many versions of this new type and method of work, too.

According to the objectives of the research, the *types of distance work* can be classified in many ways, mentioning only a few examples: according to the nature of performed tasks, the nature of communications between the employer and employee, and also according to the dependence of the employee.

- There is a basic distinction in distance work on the basis of the contents of the *performed tasks*, in which case the professional capabilities of the employee are of major importance. There is creative distance work, and there are simple activities requiring very little vocational skills. It should be noted at this point that these two

groups are generally typical, and distance work activities involving work between the two are very rare. In the first one, the employees primarily enter data without almost any vocational qualifications, while in the second one, designing and developing activities take place very often across national borders, too.

- Concerning the *communication between the employer and employee*, the frequency of personal contacts is the most important factor. In other words, distance work, in the framework of which the employee spends at least one day at the workplace each week, and has an opportunity to discuss and evaluate his/her work with his/her colleagues, and the employer personally is one category of this work. The working relationship is completely different when, in the most extreme case, the employer and employee never meet, and they do not even plan for the meeting. In such cases, the working relationship between the employee and the superior, defining and controlling his work is only maintained through the Internet. (An example for this is the virtual company, which will be mentioned later.)
- There are also significant differences according to the *employment contracts*: the employee may have all his/her income from distance work, and he/she may work as a simple employee or an independent contractor, and this type of employment could also bring supplementary income for the employee.

6.7.2. Conditions of distance work:

The *technical background,* the telecommunications devices, and adequate knowledge and skills for the use of those devices, with continuous further learning abilities and willingness are conditions for the spread of distance work, accepted by everybody. The technical requirement for distance work is a network relationship between the employer and employee, which is either established for this purpose, or exists independently from it, too. This network could be general Internet or intranet network. However, subject to the nature of the work, different requirements may occur: for example, in the case of large complicated planning tasks, there may be need to assure the periodical transfer of large quantity of data, and in the case of joint development activities of several companies, perhaps several countries, Internet communication should be made available at scheduled times. However, in the majority of cases, distance work only requires access to a network when the job is collected and the product is submitted, as well as during consultation. A very important technical requirement related to Internet activities is the protection and security, encryption of data. Generally, it can be achieved with more simple solutions on an intranet network. The spread of electronic signature, for which now there is a legislative framework, will probably support distance work from these aspects. However, the real use of electronic signatures will also require the development of instructions for execution.

The technical conditions, their availability and willingness to use them, knowledge represent issues occurring both for employees and employers. We might say that this is the basis of distance work. It is supplemented with the *human conditions,* i.e. there should be someone allocating work in this way, and on the other side, the employee should also be capable of interpreting and executing the jobs allocated this way.

One of the “attractions” of distance work is that there is no limited working time; the employee may perform the task at any time during the day. But is the “boss” also available at all necessary times? Therefore, a very important further technical and work organization issue is to organize the necessary consultations and controls between the employer and employee. This again moves distance work to the often mentioned two work areas: i.e. creative activities requiring very little consultation and strictly regulated, simple activities.

Concerning the conditions of the spread of distance work, the knowledge of relevant tax and *labour regulations* is mentioned very rarely, and the need for psychological preparation to ease tensions caused by potential “solitude” and the acceptance of the management of technical dependence are also mentioned very rarely.

In Hungary, distance work is not regulated by the Labour Code, and the Labour Code does not include the special aspects of distance work either. Therefore, it depends on the individual skills and preparation of the employees whether they are able to reflect their general and individual interests in their employment contract. This primarily affects the issues of working hours and holidays. However, it would be a very important issue to clearly regulate the problems of responsibility of non-performance due to technical reasons, and aspects related to the provision, maintenance and costs of the technical conditions of the work.

The employers have expressed their concern that the *legal regulations* related to distance work are *incomplete*. The employees do not bring up this issue, although they are struck more by this fact because their interests can only be reflected marginally due to basic organization insufficiencies, and they do not have a way to jointly represent their interests. Let us state very clearly, there are no organizations, which would represent such employees properly.

In the case of employees, let us concentrate on the Hungarian situation; even the new Labour Code does not regulate the concepts of working hours, holiday or extra work for this work method, which is so different from the traditional working method. The contractual obligations between the employer and employee are not regulated either.

The No. 1 reason for this is that at the moment, the number of employees taking on jobs in this form is insignificant in Hungary, and some of them are not even contracted by a Hungarian employer, if they have a contract at all. The second reason, as it has been indicated, is that mainly those appear at the labour market who looks for distance work, while the supply side is almost non-existent.

Therefore, Hungarian distance workers do not even mention these labour law issues, if they get a job offer at all. The situation can be characterized with the same features as the accelerated technical development, i.e. legal regulations lag significantly behind the real situation. Thus the exposure of distance workers is very significant, even if we only think of issues related to working hours. They do not have a way to insist on working 40 hours a week, and they are often forced into part-time employment contracts. In other words, compared to the work completed, their income is low. (This is mainly typical for simple work.)

With regard to this issue, it is very important to study the EU directives related to this area within the information society. The EU defined the creation of new jobs and preparation of employees as one of the key elements of building and developing an information society in the year 2000. It is true that the EU did not primarily refer to the opportunities of distance work. One of the basic issues of legal regulations is which areas should be regulated at national level, and which cases are already governed with existing EU directives, as well as aspects where law approximation is recommended.

In general, it can be concluded that typically those with high-level qualifications are capable of enforcing their interests better when creating their employment contract. These persons are less exposed to the conditions of the labour market, and they are better informed about their own rights as well. All this indicates that at the moment, there is no organization in Hungary, which would properly represent the interests of distance workers. At international level, typically non-profit organizations represent the interests of employees exposed to the conditions of distance work, as it will be described later.

The issue of *social background* is known for all those analysing and researching distance work. For the employer, it is not necessary to examine this area, and most probably Hungarian employers do not even know that it is another major issue. As it was already indicated, those working at home are not able to look after the children or perhaps sick parents more, if they intend to do their work properly. The surplus for them is that they are not limit to specific times in their work. But it is a question whether a separate

workplace can be established at home, behind closed doors, and if so, do the family members take into account this necessary requirement for distance work?

Another aspect of the same problem is that an employee who already got used to working in a community, spending a considerable time of his/her life with others, having coffee, having lunch with others, chatting on the corridor, will probably find it difficult to be alone all day, and will develop well-known symptoms of the feeling of being locked up, and symptoms of depression.

There is a group of employees to whom the previous points do not apply, and for whom this type of work could promote the process of their human development, and this category is the category of disabled people. Therefore, it would be very important in Hungary to support distance work, primarily among disabled people, because among those, the ratio of employment of disabled people with a degree is much worse than the average. Support should be given to distance work, primarily in areas where it could bring significant advantages.

6.7.3. Participants in distance work

Distance work is established when the employer and employee enter into a contract for this working method with joint consent. At the moment, there is a lot more positive fear primarily in employees in relation to this working method than the real advantages. Thus, distance work is represented at the labour market with more supply than demand, and this only strengthens the position of employers to be able to retain their special interests and reflect them in the employment contracts.

Employers:

The majority of European companies and organization accept distance work these days, yet there are still many factors hindering the wider spread of flexibly organized work processes. In a European survey, corporate decision-makers were asked about the most important barriers to distance work in their opinion. Most of them mentioned the issue of data protection in the first place. Concerns related to the quality of work and productivity were mentioned in second place. “Distance work is a good idea, but I do not think that it would work among my employees”, they said. Corporate decision-makers mentioned that the difficulty of controlling distance workers, the lack of motivation to change, high expenditure, difficulties of organization of communications, health, social security, labour problems, and, in the last place, the lack of employees’ interest in distance work, and the

resistance of trade unions were all mentioned as problems. It is worth thinking about this list containing an incredibly long list of factors. Employers use distance work primarily in order to reduce their expenditure. As a result of distance work, they can reduce their office, investment and travel expenses. If the distance works are employed part-time, or temporarily, they can further cut their expenditure.

Employees

Employees deal with the issue of distance work in two cases: if such a decision is adopted in their company on the introduction of such type of work, or if they hope for employment in this form being unemployed. It means that they mainly have extreme, and not carefully developed views, or analytical information.

6.7.4. International situation

We can assume to make a projection for distance work in Hungary on the basis of the experience of the EU countries, taking also into account the typical Hungarian features.

The European situation is represented best with the results of the surveys conducted in the territory of the European Union. In the middle of the 1990s, in 1994, there were hardly any distance workers, in 1996, there were a few millions, and in 1999 their number reached nine millions. By the end of the 20th century, by the year 2000, the number of distance workers in the territory of the European Union exceeds 10 million, according to the calculations.

These figures can really be explained, if we also take into account that in 1999,6% of the employees of the countries involved in the survey were employed as distance

Country	Number of distance workers in thousand	As a percentage of all employees
Ireland	61	4.4
U.K.	2,027	7.6
Spain	357	2.8
Italy	720	3.6
France	635	2.9
Germany	2,132	6.0
Finland	355	16.8
Sweden	594	15.2
Denmark	280	10.5
Norway	1,044	14.5

(Source: ECATT 1999, cf.www.ecatt.com)

workers, and their breakdown was also very different in the individual countries.

It is also worth taking a look at the breakdown of the 1999 survey, i.e. the 9 million employees, according to the types of distance work. There were nearly 3 million working from home regularly, there was a similar number of occasional distance workers, 2.3 million mobile distance workers (working from various premises), and there were approximately 1.5 million individual and small entrepreneurs having a small office at home.

Looking at the data, it can be seen that even according to the trends, distance work will primarily be typical in the four Northern countries. It is only partly due to the fact that the Internet use penetration is the highest in these countries. Most probably, a similarly important aspect is that the standard of life is extremely high in these countries, and therefore, there is a high number of people with a decree. However, in addition to these factors, the different human relations and habits compared to the other European countries, due to the rough weather, is at least a similarly important factor in the spread of distance work.

It is clear that the spread of distance work is due to many components. But in whose interest is the spread of distance work?

In the territory of the European Union, on average 6% of the employees complete distance work. Typically, such employees are between 30-40 years of age, they have a degree, sometimes more than one degree, and most of them are men. They do not live further than 50 kilometres from the employing company, and continuous personal contacts are also available, if necessary. Distance workers do not do more housework than non-distance workers working in similar jobs. Typically, the employment contract specifies that wages are paid according to the performance, i.e. they do not get regular wages. In a different approach, they do not have an income, if they do not have any work.

6.7.5. Hungarian situation

In Hungary, several organisations and studies have focused on the issues of distance work for many years, but the number of employees working this way is still very insignificant.

In Hungary, Telehouse movement is one of the major organisers and advertisers for distance work. The only distance work information portal in the Hungarian language is available from its homepage (with a link). In other words, it seems that there is only one

place in Hungary where one can search for such type of work through the Internet, at the http://www.tavmunkainfo.hu/ address.

These offers cover a relatively wide range of jobs from translation through accounting, all the way to graphics. Specific queries can be made with further links by both employers and employees. The offers include offers from Hungary and other countries. The portal is regularly updated. Following a study for a longer period, it can also be concluded that in Hungary those offering distance work are primarily looking for experts, and people with a university or college decree. Most probably, distance work will not reduce unemployment in Hungary significantly either.

The spread of distance work cannot even be an effective “weapon” in reducing unemployment according to those who were asked in the survey. This new working method may have a role in helping the regions lagging behind, such as for example, East Hungary region, which is developing slower, to catch up, in the process of the equalisation of regionally different economic development in Hungary. Distance work does not recognise geographic constraints. It is well-known that in Hungary, the geographic mobility of employees is very weak, and the establishment of distance work networks might dissolve this disadvantage virtually, and it may induce favourable changes. It is worth taking a look at a few Hungarian research results. In the case of questions asked about the economic aspects of distance work (in which those giving answers gave a mark between 5 and 1, giving 5 if they fully agreed with the question and 1 if they did not agree with it at all) (N=100).

The average answers and marks are summarised in the table below:

Distance work could significantly reduce the office expenditure of companies	3.55
Distance work could significantly reduce traffic, and the crowds on public transport, air pollution	3.53
Training of distance workers, and their supply with equipment would mean a significant financial difficulty for companies	3.48
Distance work might cause serious work organisation difficulty within the company (for example, in maintaining contacts, and control)	3.30
Distance work could significantly reduce unemployment in Hungary	2.49
As a result of distance work, employers would find it easier to find cheap and better qualified employees	2.49
With distance work, disadvantaged regions could catch up more easily	3.20
Distance work could significantly increase the effectiveness and performance of work	2.98

(Source: http://www.tavmunkainfo.hu/)

In the spring of the year 2000, the answers of the population given to the following questions, on a similar scale of 1-5 points, in a survey conducted on the basis of a government assignment, were the following :

Question No. 1:

Can you perform your work, and do you only have a computer and telephone contact with your workplace?

Does not agree at all	63 %
Does not agree a little bit	6 %
Agrees a little bit	9 %
Agrees partially	
Fully agrees	7 %
Does not know	11 %

Question No. 2:

Do you think that in ten years' time, the majority of people will work at home in Hungary?

Does not agree at all	29 %
Does not agree a little bit	17 %
Agrees a little bit	22 %
Agrees partially	10 %
Fully agrees	7 %
Does not know	15 %

In accordance with the experience of most European countries, and a conference organised in Budapest recently on the subject of distance work in Hungary, too, a process typical of the introduction and spread of distance work was related to the reduction of branch networks in the countryside, and the reduction of local representation offices of banks and insurance companies, as well as expensive office buildings operated by companies. A new trend is the centralisation of administrative functions in call centres. In such cases, distance work could be an instrument for having access to professional skills which are geographically distant from the head office, yet are absolutely indispensable.

6.7.6. A few specific cases of distance work

The current European situation, the interests of employers and employees, i.e. the real situation and opportunities of the participants are best described with some specific examples, using the case studies mentioned at the Telework London 2000 conference:*

- According to Alan Denbigh, chairman of the British Distance Work and Telehouse Association (TCA), the transport chaos only points to the opportunities of distance work. For those working from home, the lack of fuel and traffic jams on motorways do not represent a problem. They can continue their daily work at home undisturbed, because progress on the information superhighway is not hindered by traffic jams or strikes. Data can be shown in figures concerning the fuel and time saved by those working in distance work as they do not commute daily, and they are not affected, not even mentioning the favourable environmental impacts. (Source: Telework Conference, London 2000, Andrea Wesselényi Perfect Team Bt.)
- A detailed case by Automobile Association (AA). “AA” is the British equivalent of the “yellow angel” service known in Hungary. Because of the large number of employees, their offices were extremely crowded. 80% of the assistance calls came in during the morning and afternoon peak hours. Therefore, neither the ordinary daily eight working hours, or even shifts, did not prove to be a fortunate solution. When “AA” was restructured, 150 employees were selected from voluntary applicants for distance work, for whom it was ideal to work in their home office for 3-3 hours a day during the morning and afternoon peak period. In accordance with the project described at the conference, the advantages of the distance work programme appear mutually for the employer and employees, if the employee has a choice.
- British Telecom implemented its call centre service outsourcing the work to home offices. The results brought unexpected success. Employees working from home in 6 hours managed more calls than they did in the office for 8 hours, i.e. British Telecom managed to increase the efficiency of work and the number of employees satisfied with their conditions of life at the same time.
- The German LVM Versicherungen saved DEM 100 million by launching a distance work programme instead of the construction of the planned new office building in North Münster 30% of the employees of the head office of the insurance company, approximately 500 people work from home. LVM achieved a 10% productivity increase among distance workers.
- British Telecom, which one of the main sponsors of the conference, launched an extensive eBT distance work project, achieving to save approximately GBP 180 million property-related expenses and travel expenses related to approximately 20 million miles, as well as a 20 % increase in productivity. At the moment, more than 4,000 BT employees have registered home office, and further 36,000 employees work flexibly, i.e. not in a fixed office.
- Italia Telecom launched an experimental project last year, in the framework of which approximately 200 volunteer operators of the 2,000 employees working in the inquiry service of the telecommunications company started to do their daily work from home. The leader of the experiment explained that according to their survey, 80 % of the customer service employees working from home had a very positive opinion about work at home, and only a very small percentage of the employees faced the problems of isolation. Overcoming isolation should not be the individual’s problem, but the organisation should be responsible for it - Patrizio Di Nicola, a university teacher controlling the research underlined in his presentation.

With regard to the conference, Barry Hartrop, the Norsk Data Home Team manager, should also be quoted, who was one of the most successful presenters of the

conference, and who said that reference to statistical data was only one of the commonplaces of drawing attention to distance work. “We are all aware of the obvious advantages of distance work, have heard so many data and daring prediction for the future. However, the truth is that it is a lot more difficult than we ever thought to break up traditional types of work, introduce distance work, and have it accepted.” In accordance with Hartrop’s views, the No. 1 barrier in spreading distance work more widely is still the inflexible managerial thinking, the fear from anything new and fear of accepting new methods. “We can describe brilliant statistics, draw growing graphs about the spread of distance work, but the really good encouragement can only be given with the ‘just do it’ strategy. The more places use distance work successfully, the more accepted the flexible types of work become in more and more areas of the economy, but it is not at all a fast and simple process, as we originally thought. The breakthrough is still ahead”, - thinks Hartrop.

According to chairman Denbigh (TCA), distance work is by far not a general type of work yet, but its proportion is clearly increasing. He thinks that in the knowledge-based society and economy, approximately one-third of the employees will spend at least part of their working hours flexibly. In the UK, the number of distance workers reached 1.6 million in 2000 (8 % of the population of working age), which involved a 20 % increased compared to the previous year. In accordance with the EU data broken down according to countries, the Scandinavian countries and Holland are still in the lead in using distance work.

6.7.7. Whose interest is distance work?

It is worth coming back to this issue briefly. Recalling the case studies and the lessons of the conferences, we can conclude that the employer is the major participant in distance work. It is in the interest of the employer to reduce the operating expenditure, and not increase the wages of employees working from home at the same time, supporting only the establishment of the infrastructure required for the work. Significant financial investments, or higher wages are only required, if employers seek special professional skills and find them this way, can employ such persons only in this form.

Finally, let us quote the concept of a *virtual company* in a sentence. A virtual company is a company without any real (traditional) components, buildings or offices. Employees communicate with each other only through the Internet, and this is the way in which they develop things jointly, perform tasks in co-ordination, living on different

continents even. The most important aspects of such a company also include the economic, profit-oriented interests of the shareholders of the company.

6.8. Telecottages in Hungary

Mátyás Gáspár, author

László Z. Karvalics, reviewer

6.8.1. Briefly on Telecottages

The ever-increasing need of communal access to the facilities of the information society is granted by Telecottages. To be more precise, the term Telecottage means community teleservice centres, i.e. those public service enterprises in which users are given the opportunity of gaining access to the means of telecommunications, computers and the Internet; users are also aided in acquiring and accumulating practice in handling the whole new array of infocommunication accessories, getting to know about the services, as well as getting and applying empirical knowledge. For all those who choose not to take advantage of these services, but do need them Telecottages serves as cultural switch-plates.

It should strike one as no surprise that small-size settlements have been among the first ones to embrace the new public services in both developed and developing countries. The informational and telecommunicational infrastructure rides a wide range of possibilities as diverse as correspondence, access to information, counselling and transaction of affairs, as well as employment, business, education, promotion of culture and entertainment. As a matter of fact, it seems that each aspect of life has something to gain from Telecottages. If a small community -proportionally- is only able to allow for the establishment of one such infrastructure, then Telecottages should cater for all the needs. If schools, culture centres, libraries, the regional government, doctors, postal service providers are unable to maintain such infrastructure and if other institutions - ones promoting the development of domestic relations, entrepreneurial aims, banking services - are still lacking in the area of a small settlement, then all these governmental, civil and business necessities can jointly be overtaken by Telecottages. This is the exact state of affairs in Hungary.

Telecottages can succinctly be described as public service enterprises equipped with modern informational and communicational technology whose main objective is to allow small-size communities to gain access to the infrastructure required for whatever

reasons deemed necessary. Telecottage are institutions drawing on a more limited scale of technical and service-wise background and can even represent a new telecottage in evolution. In is also common that telecottages rely on services offered by larger enterprises in their vicinity in order to make the range of services more extensive. In palpably small-size settlements in which no more than a hundred people live there is often no need for an all-encompassing set of telecottage-services (i.e. there may exist no demand for the advancement of socially disadvantaged regions).

6.8.2. A Hungarian success story (?)

The Hungarian Telecottage programme has been regarded as a success story both at home and abroad. In May 1994 in the library of Nagymagócs (Csongrád County) and in a restored cottage[23] in Csákberény in June the first rural telecentres were established. The Hungarian Telecottage Association, established in 1995, agreed to the terms of the National Telecottage Strategy in which the following forecast was made regarding the development of Telecottages in 2000 (it appears that this prognosis has worked disregarding the minor negative fluctuation in the development of Telecottages). The expansion of the movement seems to be on the right track given that the national scheme for development known as Széchenyi Plan has accepted the movement and granted its further advancement and expansion.

	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003
Number of Telecottages	2	2	5	20	80	150	200	400	600	800
Number of telecottages	-	-	-	-	20	50	100	300	600	1200

There has accumulated ample material both on the Internet (webpage, multimedia magazine, CD) and in a more traditional form (articles, newsletters, books, video) on the topic of Telecottages; as a consequence, two questions must be answered as regards the Telecottage movement: (i) what is the key to this success and (ii) is there place for Telecottages in the future.

The Telecottage movement was started as an excellent initiative which has proved to entice interest and enthusiasm; it seems that its network cannot be stopped from expanding further. Telecottages jointly provide the public, civil and business spheres of interest with a wide array of services, more than 60 altogether. The possibilities are

[23] The author in tandem with his family is the founder and the promoter of the Telecottage in Csákberény, as well as the founding president of the Hungarian Telecottage Association Union (1995) and the first director of the Telehose Non-Profit Company (1999).

gradually explored and put to application based on regional characteristics, which explains the reason for the qualitative difference in services provided.

There is, however, a tangible problem: if Telecottages fail to abide by the norms of uniform quality assurance, the introduction of the monitoring managerial system as well as the engagement in the network system (electronic commerce, distance work, distance administration, distance learning), the possibility of doing so will soon be lost. A number of those Telecottages which have built upon the notion of civil or budgetary considerations are still behind with the exploitation of the network and entrepreneurial spirit. The time available for overcoming these hurdles is short. Local 'missionaries' are faced with the difficulties of finding the appropriate strategy for the implementation of Telecottages on a local level, which seems to call for a newly devised plan on a settlement-to-settlement, community-to-community basis.

6.8.3. The Hungarian model

The forms of communal access known as multipurpose (multifunctional) teleservice centres (or simply telecenters) can take various forms around the globe. In general, public access to the Internet is referred to as e-gateway. Depending on the range of services provided (be they civil, state-owned or business), as well as on the sphere of activity within which they operate, there exists a whole array of different models (more than a dozen different ones) as diverse as telecenters (electronic mail), Internet coffee houses and communal telecommunication centres. The multi-functional deployment of the informational and telecommunicational infrastructure within the confines of one institution reaches the best possible results in the Hungarian model in which modern technology combines with communal solidarity, self-organisation and self-management of small communities. What lies at the heart of the success story of the Hungarian model?

In the *first place*, Telecottages are organised out of self-determination with the deployment of local initiative with the aim to satisfy local needs. Therefore, in this sense, this can be labelled a 'movement from the ground up' in traditional domestic terms. The western terminology employs the following term: 'movement from the roots'. The crux of the matter is that the basis for this kind of organisation is provided by local activity and local initiative, which also, needles to say, means bearing full-fledged responsibility vis-a-vis the community.

Secondly, the majority of Telecottages was reliant upon external (foreign or domestic, mostly governmental and in recent times business) sponsorship (e.g.

computers, software) and this is the trend envisaged for the future. As can be seen, the financial hazards of the innovation are not directly imposed upon local communities or only to a lesser degree. This is both necessary given that the target of the movement are socially disadvantaged regions, settlements and small-size communities and it is also reasonable since establishing network end-points is not merely an end in itself.

Thirdly, the establishment and subsequent fate of a Telecottage becomes a communal issue which is embedded in the local society and control. Thus, in the majority of instances competitions relegate the maintenance of Telecottages to local civil organisations. Although this type of co-operation is not always smooth (institutions of local government would also like to see Telecottages under their purview), in the long run the self-organising, self-managing aspect of local communities as well as local capital are reinforced in this manner.

Fourthly, the range of services offered by Telecottages is characterised by a multitude of different sectors as well as a high degree openness towards new proposals. Telecottages, it seems, cannot fully be characterised by any field of activity, any sector of social performance. This can sometimes lead to a crisis in identity: what set of social values will Telecottages be characterised by (communal, business, etc.). If co-workers are not flexible enough and do not embrace new initiatives, the problems of this kind will only be amplified; however, having a multitude of sectors, provided it can be handled professionally, is a great advantage which can ultimately lead to proportional exploitation of possibilities and can also lead to an increase in fund-raising.

In the fifth place, Telecottages can be considered to be hybrid in nature as far as their organisational background is considered. The long-standing practice has it that the recipient organisation is usually a non-profit organisation, the body of the local government, an entrepreneur or a company. Other means of co-operation also exist but the operation of a Telecottage is characterised by bi- or multilateral contracts, which results in a ‘joint effort’ enterprise. This enables Telecottages to be multifunctional in view of the needs of local infrastructure and demands.

In the sixth place, Telecottages serve the purpose of making the public alert towards the new developments in information society. In this manner communal interest is sharpened in the direction of the new developments taking place in the world. Telehuses capitalise on the power exerted by the Internet, computers, the overall propaganda of the information society, its commercials and promises.

6.8.4. Future of Telecottages in general

The question arises naturally whether in the long run this means of 'forced' communal access has any prospect given the ever-increasing rise in individual access to the information society.

A positive answer can be supported by twelve arguments. Communal access and Telecottages do have a future not to be counterbalanced by individual access but rather as a mutual complement to it. The function of Telecottages forming a bridge over the hiatus on the digital plane in underdeveloped and developing countries vis-a-vis the developed countries is obvious, but it seems that developed countries can also benefit from it.

1. Income conditions. Unfortunately, even in the near future there will be low-income individuals who do need the means and services of infocommunication but are unable to cater for the conditions of continuous access on their own. Public institutions (e.g. schools, libraries, cultural centres, etc.) are unable to grant this function of communal access or are able to do so only to a limited extent.

2. Cultural distance. Overcoming the distance between the traditional and the new telecommunicational culture calls for external help. Telecottages play a prominent role in this. They translate the needs formulated within the confines of the old culture into a new cultural vocabulary. This function is not transient in nature as there is an ever-increasing step with which new technologies (e.g. the new mobile culture) are being developed.

3. Technical novelties. New accessories, especially new software, need to be tested before they are allowed to be disseminated in wider circles. The manufacturer and inventor of such new means require instantaneous feedback, which can be handled most productively in such communal access centres.

4. IT assortment and supply. Communal access can also be conceived of as a kind of demonstration centres, which is mostly of interest to the business sector. Products and services must be introduced to the widest possible audience. Can there be a better solution to this than a communal access centre in which prompt business deals are called into being.

5. Rational utilisation. There have always been various means whose exploitation is more sensible in a communal centre than individually. The main consideration behind this is not solely that of financial appropriateness but rather full-scale utilisation and keeping pace with technical depreciation. If someone is only occasionally obliged to use scanners, colour copy machines and printers, video-conference utilities, etc., why should such appliance be kept at home?

6. Personal innovation hazard. Aid is needed for those individuals who choose among new accessory and services to ascertain whether they can be successfully utilised in an activity, whether they are costworthy and time effective. Communal access can have a beneficial effects on attenuating personal innovation hazards.

7. Social innovation hazard. The information society has produced a number of changes in distance learning, distance work, a number other distance activities and of course in electronic democracy. Communal access centres can serves as small-scale test environments for ascertaining the effectiveness of certain projects

since these small-case enterprises can be organised more easily and their soundness gauged in a less hazardous environment.

8. Community control. The information society contains inherent dangers as well as advantages. The communal society aided by public control can find a way to eliminate and attenuate these problematic areas. In Telecottages one cannot be affected by deleterious effects, the users of services can be institutionally safeguarded, the emergence of dependence can be prevented and the interests of small communities and the people therein can be kept a close eye on.

9. Community needs. Telecottages, in addition to satisfying personal needs, also offer the possibility of communal utilisation of informational and telecommunicational technologies. Communal utilisation of informational and telecommunicational possibilities is as diverse as the construction of local websites, operation of local radio and television studios, organisation of local publicity (e.g. newspapers), attending to the needs of local economic development, etc.; all this is attempted with the use of modern technologies and network access.

10. Global needs. Public access to the network ids a global need. Not everyone is expected to be equipped with mobile computers. Irrespective of the pace with which the informational society is gaining ground, visitors and travellers will always need public access to computers, the network and each new technological commodity to appear in the future.

11. Security requirements. Telecottage commodities and informational database can be guarded against the threats coming from informational crime in a more effective way than it is possible from computational appliances operated on domestic grounds. The personal, technical and other requirements can be supplied on communal grounds in a more cost effective way.

12. Fund-raising and maintenance. It is easier for a community coupled with individual co-operation to gain access to funds; the reason for this lies partially in the fact that some resources are only accessible to a community as a whole but also in the fact that *spatial requirements* can only be guaranteed on a communal basis.

In the course of writing these lines, the following news has arrived: in Great Britain 1200 Telecottages have been inaugurated in addition to the already existing 200 ones; in the years to come altogether 6000 new communal access centres will be erected. Similar programmes are being established in Canada. The developing countries of Africa, Asia and Latin-America are already known for such enterprises. This appears to be a global process. Is it possible that Telecottages are the forerunners of a developing wave of public information networks?

6.8.5. The future of Telecottages - here and now

The two perspectives of *long-term sustainability*:

Shouldering of communal duties and the public network. Telecottages from the moment of their conception and in the course of their functioning have always met the

demands of communities. Civil non-profit organisations have understood the notion that wholesale prosperity of communities is largely dependent on computers and the network in this information society of ours. The bodies of local self-government and the central government grasped the social importance of the Telecottage movement as far back as 1993 and have ever since attempted to provide ever-growing financial means for its expansion. The further course of this process can be given the following forecast:

- The basic set of Telecottage services, i.e. public access to the network and use of computers, in the form of communal duties which are already accessible in some public institutions (especially in libraries, culture centres, schools, etc.) and civil organisations, could be promoted to a budget-financed nation-wide project bound by a contract for public duties with the proviso that Telecottages be responsible to maintain a basic level of services in the future.
- 15 – 20 specific public tasks appeared in the practice of Telecottages, from administration and teaching through social services and environmental protection to local economic development and service for village economists and region developing managers. The definition of their Telecottage-techniques and basic services could provide a basis for the supplementary normative aids, for contractual duties and wide spread of those activities in the network.
- The wider the circle is where Telecottages provide public services, the more justified it is to establish an aim-oriented public network of rented telephone lines and to finance it as part of the normative funds, similar to the Irisz-Sulinet network system. Telecottages that provide public access to basic services and other institutions could also be given the opportunity through which substantial benefitions and funds could be provided to disadvantaged persons and families in order to eradicate the informational hiatus.

Telecottages in the economic network. The local economic development demands the collaboration of the community and the control of the society, as well as the use of modern information and communication technologies. Telecottages are able to connect the small region, the settlement and the individuals (employees, entrepreneurs and people who sell and who buy) into the electronical economy. The telecottage that provides public accession, teaching, incubation and other services can use its income for its operation and development.

The following conditions are necessary for the stabile and generally capable role of informational basic network: The continuously expanding Telecottages which are in the state of being institutionalised require additional conditions in order to become generally and steadily well-equipped for the role of basic informational network.

- The network can develop significantly in providing state and economic tasks and services, if the *network-development is predictable* and its *program will be elaborated country-wide*, while the program itself is *guaranteed by certain financial sources*. A perspective of this kind appears to be substantialized comprehensively

through the Széchenyi Plan and the respective sectoral conceptions, for instance in the field of social, rural developmental and advancement of public administration.[24]

- A *Global Quality Insurance and Monitoring System for Telecottages* should be implemented, which can guarantee that all organizations under the name of 'Telecottages' and which are chategorized appropriately, answer the same requirements anywhere in the country and, they continously provide the same basic services. Telecottages can count on receiving public duties and normative funds under only these conditions. This task is very well-described in the National Telecottage Strategy and its elabouration can start, in optimal case, already in 2001.
- There is a need for organizing *Manager Training and Personnel Qualifying in the Telecottages.* At present, approximately five hundred people work in the Telecottage Network. After the entire network have been built up, the number of "Telecottage People'' will raise to a few thousands. We need well-trained colleagues for the guaranteed service quality. The development of trainings, which give certificate for the workers after the course, are in process. The task is quite unique and individual, as Global Telecottage Trainings are not organized even in the most developed countries, like Great Britain, either.
- There should be a so-called *Telecottage Mentor & Monitor Network* organized on regional basis, in order to provide assistance in the establishment of new Telecottages and introduce of new services. Its members, who are well equipped with theoretical, methodical and practical knowledge, should be able to provide assistance on the spot, to collect experiences and transfer those and to evaluate the local circumstances (for instance, for Quality Insurance). The Telecottage Movement already has the appropriate people for these tasks.

6.8.6. What kind of Telecottages?

Defining the word "Telecottages" always resulted difficulties and still generates discussions. The first problem to face is that Telecottages can be defined based upon their service capacity (infrastructure) and operation method and not based upon their activities. The suspicion that Telecottages steal functions from already existing institutions (especially from libraries), is understandable. But would libraries be able to provide all services and take upon all tasks in the same way and as open as Telecottages do?

[24] The opinion of the educational sphere is very interesting and contradictory. Meanwhile one third of Telecottages operate in institutions that give significant supplementary funds for the development of the section, the representatives of the institutions behave inimically with the movement, despite the many good examples and cooperating partners. They say that there is no need for Telecottages, as the cultural houses and libraries of the villages can provide the same duties. They think that the movement takes all financial sources away from them. The question is, however, who could prohibit the employees of village libraries and cultural houses from establishing a Telecottage and who could benefit from the opportunities that are provided by the Telecottage Movement? Some difficulties could be, however, that Telecottages operate mainly on civil basis. I talked about its necessity at the beginning of my study. The employees of educational institutions are civil persons at the same time, why do people think that they cannot stand on the front of the local civil organizations, as some of their colleagues do it at certain places? Telecottages resulted the rebirth of educational institutions and certain libraries. Should we really kill the movement, just beacuse the section would require a task that was not provided by itself? Would it be the modern solution for the equality of spheres, ensuring free competition and strengthen civil self-initiations and voluntarism?

Nevertheless, despite the different characteristics, the different institutional supplies and the small community's culture, the size of a settlement and its public administrational rank is not really important in the use of the name 'Telecottages'. The most important factor is the implementation of the basic-telecottages. The need for a Telecottage can be found even in a city when the conditions for establishment and operation of a Telecottage are ready. The stipulation of basic-telecottages means that the circle of the users cannot be closed and limited, thus, it has to provide services to anyone who needs that kind of infrastructure in a settlement or in a regional community. The definition of such a Telecottage can contain, of course, that the majority of the users are from the circle of children, romas, pensioners or handicapped people. This, however, does not mean that other layers of the society are excluded from the use of the Telecottage. In this way, it is even suggested to establish and operate such Telecottages, as they can contribute to the fulfillment of specific needs of certain target groups, which services afterwards could spread around in the network and provide assistance to the target groups even there, where there are fewer representatives.

Literature

[6.8.1] Átjáró, Nemzeti Teleház Stratégia és Program 2000-2006. Összefoglaló, Szerkesztette: Gáspár Mátyás, Teleház Kht, 2000. (Passageway, National Telecottage Strategy and Program)

[6.8.2]. Perspectives on Distance Education Telecentres: Case Studies and Key Issues, Szerkesztették: Latchem, Colin, Walker, David, The Commonwealth of Learning, 2001., Vancouver, www.col.org

[6.8.3] Community Teleservice Centres, CTSC International, 1994 (A nemzetközi teleház szövetség jelentése a világ teleházairól).

[6.8.4] Diamond, David, Elektronikus falu Virginiában, Reader's Digest, 1997. február (Electronic Village in Virgina)

[6.8.5] Erdősi Ferenc dr., Telematika, Távközlési Kiadó, 1992. Benne: A falusi települések fejlesztése "teleházak" segítségével (Developing Rural Settlements by Telehouses)

[6.8.6] Gáspár Mátyás, Jövőtervezés Csákberényben, A falu, 1994. tavasz (Future Planning in Csákberény)

[6.8.7] Gáspár Mátyás, Közösségi információs és szolgáltató központok működésének tapasztalatai az USA-ban, Főszi dokumentumok, 33. 1985. július (Community Information and Service Centres in the USA)

[6.8.8] Gáspár Mátyás, Takáts Mária, Építsünk teleházat! Erdei Iskola, Magyar Teleház Szövetség, 1997. (Let's Build Telehouses! with contents in English)

[6.8.9] Gáspár Mátyás, Teleházak Magyarországon, Mi a jövő? Szerk.: Bognár Vilmos, Fehér Zsuzsa, Varga Csaba, OMFB, ORTT, HÉA Stratégiai Kutató Intézet, 1998. (Telecottages in Hungary, in: What is the Future?)

[6.8.10] Heinz Dörr, János Kárász, Johann Steszgal, Telematik für den ländlichen Raum, 1990 Wien

[6.8.11] Hopkins, Cleve, Community information and services centres: Concepts for activation, U.S. Department of Commerce, July, 1976.

[6.8.12] Könyvtárak és információs intézmények szerepe a demokratikus tájékoztatás szolgálatában, Országos Széchenyi Könyvtár Könyvtártudományi Módszertani Központ, 1993. (The role of libraries and information provider institutions in the democratic information service)

[6.8.13] Közösségi élet, közösségi szolgáltatás: teleház, Magyar Távközlés, 1998/2. (Community life, community service: telecottage)

[6.8.14] Lars Qvortrup (ed), Learning at a distance, CTSC International Newsletter, Volume 3, 1992.

[6.8.15] Lars Qvortrup, Community TeleService Centres: A means to social, cultural and economic development of rural communities and low income urban settlements, 1994.

[6.8.16] Lilian Holloway, Telecottages, teleworking and telelearning, Teldok report, Stockholm, 1994.

[6.8.17] Magyarország az információs társadalomban, Szerkesztették: Fehér Zsuzsa, Bognár Vilmos, Varga Csaba , ORTT-OMFB, 1998. (Hungary in the Information Society)

[6.8.18] Nemzeti Információs Társadalom Stratégia I.0., IKB, 2001. május (National Information Society Strategy)

[6.8.19] Sixtus Lanner dr. (szerk.), Die Welt im Dorf, Telematik im Landlichen Raum, 1990.

[6.8.20] Schuler, Douglas, New Community Network, Wired for Change, Addison-Wesley Co. 1996.

[6.8.21] Széchenyi Terv, Információs Társadalom és Gazdaságfejlesztési Program, IKB, 2001. (Széchenyi Plan, Iformation Society and Economic Development Program)

[6.8.22] Telecottages in Hungary, The Experience and the Oppurunities, Szerkesztette: Bihari Gábor és dr. Jókay Károly, IGE, 1999.

[6.8.23] Teleházak és távmunka Magyarországon, Szerkesztette: Gáspár Mátyás, Teleház Kht, 1999. (Telecottages and Distance Work in Hungary)

[6.8.24] "Telekunyhók a vidék fejlesztéséért", Országos Széchenyi Könyvtár Könyvtártudományi és Módszertani Központ, 1992. (Telecottages for the Rural Development)

7. Operation

Introduction

In this chapter there are discussed issues, which influence to a considerable extent the operation of the telecommunication networks and - from the perspective of the user of the network - are determining the quality of the network.

For all telecommunications services the time ratio in which a service is available for the user with the stipulated quality is an important feature. The continuous service usage and availability of the systems that realise the services could be obstructed or made impossible by network failures. The user's reaction, if a service is not available in a given time, can be different. Beside the bad perception about the service, in some cases the users may suffer particular financial or economic loss due to the lack of (or in quality degraded) telecom services. The Service Level Agreements (SLA-s) usually prescribe those tariff and compensation conditions according to which the service provider shall compensate the user in case of a service outage. Therefore, it is recommended for the telecom provider to plan the availability level of the service in advance. Availability planning always means considering of several aspects and can be carried out only by probability calculations. It is the consequence of the nature of events affecting the availability is random, but the processes can be properly modelled by mathematical tools. After all, the reliability analysis is a preliminary step before the economical decision on to what extent it is worth to invest money in availability improving technologies and/or enhancing the operation and maintenance conditions.

In the last decade more and more services are added to the traditional telephone services and the demand for transfer of high bandwidth data is increasing. The ability to deliver voice, video and data at higher speeds is becoming a critical requirement. Now, the telecommunications networks have to integrate different computer networks with different bandwidth demands. Their different nature of the traffic also generates new requirements for Network Operators and Service Providers.

In order to be able to carry different kinds of data at a faster rate, quite new technologies and significant improvements to the existing technologies and protocols have been and still are being introduced. The complex and diverse equipment and service portfolio, their rapid change makes the planning, installing and maintaining of the controlling networks, very sophisticated. This problem can be solved only by proper and adequate automation.

Telecommunications Network Operating Organisations have invested significant amount of time, money and human resources into building up their complex telecom networks that need to be maintained, as well. In order to maximise its efficiency and productivity, the telecom network is to be controlled and managed from the following three points: network elements, services and traffic.

Proprietary (vendor-specific) network management cannot provide interoperability among diverse technologies, equipment and networks deployed by different manufacturers and proprietary management solutions having different features and different level of management capabilities. Theoretically, the gap can be over-bridged by human resources, however, the speed that can be provided in this way is often not fast enough to provide the contracted level of service for the customer.

As the communications industry continues to evolve with deregulation and liberalisation, service providers are under increased pressure to deliver a broadened set of services at a competitive price. As a result, service providers must deliver these services in a cost-efficient and timely manner. The ability to manage effectively these networks has become a key point in retaining existing, as well as acquiring new market share. However, the complexities surrounding today's networks present challenges in achieving of the effective network management goal.

Traffic management in telecom networks deals with the controlled use of network resources to prevent the network from "having a bottleneck". In particular, when more traffic should be transmitted on the network that they can effectively handle, network performance will be degrades. Traffic management controls traffic generated by calls entering in and flowing through the network, and prevents the network from overload.

Telecommunication installations and equipment accommodated in them may vary by their design and destination, but they are common in one: all of them inevitably need electric power. In most cases, the public electricity network supplies this energy. In some particular cases it happens that the electric power supply is not available for the telecommunication system. Here, according to the local conditions, alternative energy sources provide for the feeding of such installations. Due to the fact that in such cases the energy is discontinuously available, these systems are supplemented with interim auxiliary energy storing batteries. In Hungary – with the exception of some special cases – the electric energy is available all over the country.

However, the continuity of the electric power supply cannot be always ensured, for a shorter or longer time interruptions (failures) may occur due to the outage of the mains, or due to the transient phenomena, disturbing the operation of the powered equipment. Telecommunications equipment must work also under such circumstances. During the powering of the telecommunications equipment with energy, the main goal is to provide for the possible most safety and reliable feeding; therefore it is necessary also the generation and storage of the energy in a telecom building.

In the interest of proper functioning of equipment in the given environment, i.e. without the degradation of its performance, it is necessary that the different equipment and their operational environment are made electromagnetic compatible. *The electromagnetic compatibility* (EMC) refers to such a professional field, *the object of which is to eliminate or at least to possibly minimise the "mismatch" between the equipment and their operational environment* according to the accepted norms, standards and regulations.

For the selection and implementation of a telecommunications system there are quite a lot financial and technical conditions to be taken into account. The first task of the planner/designer is to determine and specify the major functions of the system, i.e. to decide what is needed. This usually includes a technical analysis, a system planning, based on marketing survey. The Buyer issues a tender invitation on the basis of the large-scale and later the detailed system plan. By the thorough technical and financial assessment and evaluation of the received bids the most appropriate telecommunications system can be chosen. The system selection

means, therefore, such a technical/financial analysing and planning process, in the course of which the Buyer – with regard to the market conditions – brings into harmony his needs with his possibilities.

Csaba Kántor dr., Editor of the Chapter

7.1. Factors influencing the availability

Géza Paksy, author

József Wiener, reviewer

7.1.1. Importance of network availability

For all telecommunications services the time ratio in which a service is available for the user with the stipulated quality is an important feature. The continuous service usage and availability of the systems that realise the services could be obstructed or made impossible by network failures.

The user's reaction, if a service is not available, can be different. Private users may have bad perception about the service or can be disappointed. On the other hand, corporate users may suffer particular financial or economic loss due to the lack of telecom services. The service level agreement (SLA) usually prescribes the way and extent of compensation that the service provider shall make in case of service outages. Therefore it is recommended for the provider to plan the availability level of the service in advance. Availability planning always means considering several aspects and can be carried out only by probability calculations. It is because the nature of events affecting the availability is random, but the processes can be properly modelled by mathematical tools.

After all, the reliability analysis is a preliminary step before the economical decision on to what extent it is worth investing money in availability improving technologies and/or enhancing the operation and maintenance conditions.

7.1.2. Availability definitions

The quantitative measure of availability of telecommunication systems and networks is the *Availability* defined as follows:

$$Availability\ (A) = \frac{error-\ free\ operational\ time}{total\ observation\ time}$$

The availability is usually given for a year, or rarely for a month.

The un-availability rate (UN) can be calculated as follows:

$$\mu A = 1 - A$$

Th availability of professional telecom systems falls into the range of 99.5…99.99%.

The availability of a telecom system depends on several factors that determine the reliability of the entire system. Since the system reliability depends on these factors, they are together called “dependability”. Although the dependability itself has no measure, the parameters introduced in the following part have.

Reliability

The *Reliability* (R) parameter gives the probability that a device or equipment does not fail during a time period of *t*. In case of standard deviation, in addition to the expected reliability value, the reliability is also featured by the Mean Time Between Failures (MTBF) parameter.

The *failure rate* (λ) gives the average number of failures for a time unit. Its unit is FIT (Failure In Time):

$$1\,FIT = 1\ failure / 10^9\ hours$$

The relation between MTBF and λ is the following:

$$MTBF = \frac{1}{\lambda}$$

The effect of repair on availability

The *repair time* gives the mean time that is needed for eliminating a failure. This parameter is called MTTR (Mean Time To Repair).

The *repair rate* (μ) determines how many repairs fall into a certain time period (hour, month, year) in average.

Based on the parameters described above, the availability can be calculated as follows:

$$Availability\ (A) = \frac{MTBF}{MTBF + MTTR}$$

7.1.3. Effects of network failures on services

The service outages due to telecom network element failures have different effects depending upon the duration and type of the service. Short outages (some seconds) in line switched networks (e.g. PSTN) causes disconnections, but in packet switched networks it causes only some packet losses. Much longer outages (10-20 minutes) may result more serious problems in everyday life, sometimes indignation of the people may arouse (e.g. unavailability of cache machines, interruption in TV broadcasting, etc.).

The effects of various network outage times are summarised in the table below, giving a support to network planners for the network reliability calculations.

Most of the outages last 2 sec… 5 minutes. This is the domain where the service layers network management systems should act to repair the connections by rerouting or activating the protection systems.

Outage duration	Effect on service and users
< 50 msec	Minor quality degradation, certain fast line protections are activated; Audible clicks due to synchronisation losses in PSTN connections; Significant increase in bit error rate of data transmission;
50…200 msec	Some PSTN connections are disconnected. Some packets have to be re-sent in X.25 and TCP/IP systems;
200 msec…2 sec	Several PSTN connections are disconnected; Sufficient packet re-sending is required;
2 … 10 sec	All PSTN connections are disconnected and it is impossible to re-establish them; X.25 and ATM connections are disconnected, disturbances occur in data link layer, re-routing functions are activated;
10 sec … 5 minutes	All data transmission connections are disconnected;
5 minutes … 30 minutes	Disturbances and congestion may occur in the traffic, causing slight effects on everyday private and corporate life;
> 30 minutes	Major network outages affecting numerous users; If the fault concerns the important telecom operators, it may have an impact on everyday life and business; The event becomes a news.

Table 7.1.1

7.1.4. Factors directly effecting network availability

Equipment reliability

The reliability of telecom equipment is determined by its individual active and passive components. These can be as follows:

- semiconductor devices, discrete or integrated circuit elements;
- passive electronic devises, such as resistors, capacitors, inductors, etc.;
- printed circuit boards, internal cablings, rack and subrack cablings, printed circuit back-planes;
- active and passive optical devices (e.g. optical transmitter and receiver diodes, WDM filters, optical switches, power splitters);
- dismantleable electrical and optical connectors;
- non-dismantleable electrical and optical connectors.

The manufacturers determine the reliability values of the individual elements by measurements. The figures rarely mean the probability of a catastrophic fault, rather the probability of being out of the specification range. In practice, the probability of an electronic device fault (λ) falls into the range of (10…1000) x 10^{-9} = 10…1000 FIT.

The failure rate of the equipment, cards, units and ports of which the given configuration is made up, can be estimated by summing up the failure rates of their individual system components:

$$\lambda_s = \pi_e \cdot \sum_{i=1}^{N} \lambda_{si}$$

where:

N: is the number of system components;

λ_{si}: is the failure rate of the individual system component;

π_e: is the environmental failure rate multiplicative factor.

The manufacturers give the calculated λ_s figures for the delivered equipment, which are the inputs for the detailed reliability analysis of telecom networks.

Environmental conditions

The lifetime and failure rate of the telecom equipment is determined by the environmental conditions such as operating temperature and relative humidity. The current silicon based electronics technology requires +5…+40 ^{O}C temperature range and 80…85% humidity. The manufacturers have to ensure by serious planning that the temperature of the single components inside the equipment does not exceed the specified value. This can be achieved either by passive or active forced cooling.

Overvoltage and lightning protection

The inducted currents due to shortcuts of high voltage networks or power lines near to telecom equipment and cables are not allowed to cause permanent faults. Telecom networks must not fail if an atmospherical discharge occurs not directly through the network. The protection against such effects has to be thoroughly planned. (See details in Clause 7.8.)

Mechanical stresses

Telecom equipment and cables are exposed to considerable mechanical stresses. The equipment should be resistant against shaking. This is especially important in out-door equipment near to heavy traffic roads.

Probably the most frequent failure event is the cut of telecom cables due to external stresses. The failure rates can be decreased by clear signs in outer areas indicating the exact cable route, and by accurate registering of public utilities based on geospacious information systems (GIS).

7.1.5. Influence of transmission performance parameters on availability

The temporary or permanent degradation of performance parameters of the transmission routes, or a certain sections of it, interconnecting the service nodes, can make the whole connection unavailable. Insignificant transmission performance degradation is not considered to be a fault.

According to the ITU-T, a connection is considered unavailable if the transmission quality is severely decreased during 10 consecutive second. The

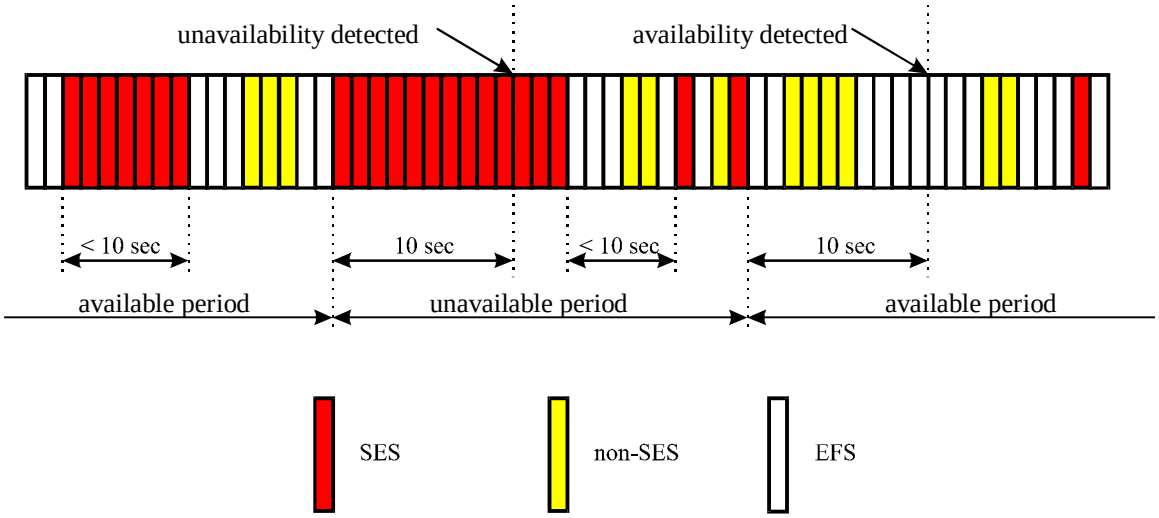


Figure7.1.1

definition of severely errored statuses is different for various technologies and bitrates. For 64 kbit/s digital connections, the G.821, and for higher bitrates, the G.826 Recommendations are relevant. The determination of unavailability is depicted in figure 7.1.1.

The unavailable periods, determined by the above showed method, shall be added to the total accumulated system outages of the system. In order to minimise the number of the unavailable periods, the outage intensity is also specified. Please refer to the ETSI standard given in Bibliography.

Translated by Róbert A Horváth

7.2. Methods to increase reliability

Géza Paksy, author

József Wiener, reviewer

7.2.1. Network reliability analysis

The most suitable method for analysing the reliability level of a telecom network is based on mathematical modelling which allows us to determine the availability of the network relatively exactly. Such model is the continuous-time, homogeneous Markov chain. In this approach, the reliability model is given by a block diagram and a state-transition diagram, determining the relation between network elements and the reliability states of the network elements. For example, the figure below shows the model of a parallel, redundant optical network. The possible states of the system are 1, 2, 3 and 4. The state-transitions are indicated with lines interconnecting the different states. The system goes from faultless State 1 to State 2 or State 3 with the probability of λ_1 or λ_2 respectively. The system is available in State 2 or 3. In State 4 the system in unavailable, and the probability of this is $\lambda_1 \cdot \lambda_2$.

If a system contains *K* sub-systems, and each sub-system has *m* pieces of replaceable or on-site reparable components, then the Markov model has $2^{K \cdot m}$ states. In most cases, the failure rates of the components are orders of magnitude smaller than component failure detection, failure recovery and repair rates, therefore the model can be truncated to fewer states.

If the repair rate of the *i*th sub-system is μ_i, then the μ_i will appear on the relevant line.

P_j probability of state *j* can be calculated from the time the system spends in state *j*, provided $\Sigma P_j=1$ and $j=1, 2, \ldots K$. The P_j state probabilities are given as the result of a K element linear algebraic system. The state probabilities of simultaneous faults in large systems are so small that they can be neglected in order to reduce the computing time.

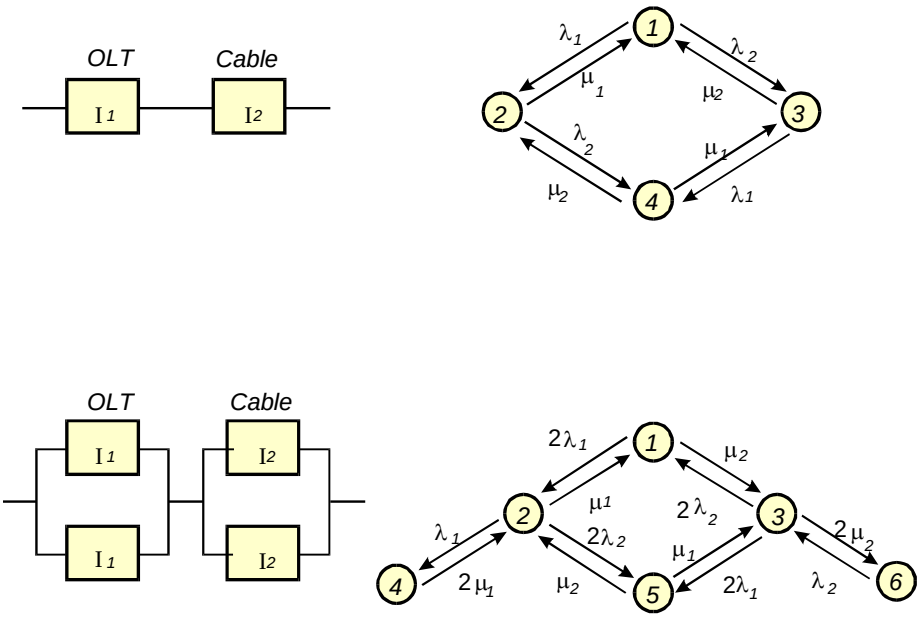


Figure7.2.1

7.2.2. Network reliability level from economical point of view

The expected continuous availability of telecom services and the level of network availability are based on the agreement between the service provider and the user. For the service provider, the availability is not only a performance parameter but an important economical factor as well, since the increase of reliability is possible only with certain investments and additional operational costs. For profitable business, the network operator should make proper balance between expenses and incomes, that is:

$$\max\{LCR - LCC\},$$

where:

LCR: is the income during the network lifetime;

LCC: is the expenses during the lifetime.

The largest profit can be gained by minimising the present value of network:

$$\min\left\{C_I + \sum_i \left(C_{ti} + C_{mi}\right)\right\} \cdot d_i,$$

where:

C_I: is the invested money necessary for a certain level of network reliability

C_{mi}: is the operational cost in the *i*th year

C_{ti}: is the lost of income due to network faults

d_i: discount rate in the *i*th year

C_t is the total loss due to network unavailabilities, directly influencing the income. This method gives us an unbiased figure for a certain availability level, rather than considering subjectively different availability values. Further advantage is that it is independent form the applied technology, and well suited for preparing different network plans.

The most critical part of the method is the estimation of C_t loss of incomes. Here is a simplified model for this:

Different effects of traffic outages on line switched and packet switched networks

In case of line switched traffic outage, the loss of income is:

$$C_t = \lambda_F \cdot MDT \cdot \alpha \cdot A \cdot E \cdot C_S ,$$

where:

λ_F: is the fault rate (fault/year);

MDT: is the Mean Down Time (hour/year);

A: is the traffic in business hours (erlang);

α: is the ration of business hours and outage time;

E: is the probability of traffic congestion;

C_S: is the service income (income/erlang hour).

In case of packet switched traffic outage, the loss of income is:

$$C_{tp} = \lambda_F \cdot MDT \cdot \gamma \cdot \beta \cdot \varepsilon \cdot C_p ,$$

where:

γ: is the packet transfer in business hours (packet/hour);

β: is the ration of packet transport during outage time and business hours;

ε: is the probability of packet loss during outage time;

C_p: is the economical consequence of a loss of packet (income/packet).

In case of leased line traffic outage, the loss of income is:

$C_{tl} = \lambda_F \cdot MDT \cdot n \cdot C_L$,

where:

n: is the number of leased lines;

C_L: is the income from leased lines (income/erlang hour).

If a network segment is unavailable, the losses of income for the services operating on this network segment has to be added together.

The parameters in this model are presumed to be statistically independent from each others, and do not change in time. However, in reality, there may be some correlation between the business hours and the fault time.

After these calculations and the evaluation of the results, the required network reliability can be ensured by the measures considered to be economic.

7.2.3. Allocation of availability parameters to international connections

The resultant availability of long distance connections has to be provided jointly by the service providers that operate its sections. So that the operators use links with proper quality, ETSI elaborated a standard to allocate the availability parameters (EN 300 416). A 2500 km long leased line reference connection is determined as shown in the figure below:

PEP: Path End Point

TIC: Terminal International Centre

FS: Frontier Station

IB: International Border

ICPCE: Inter-Country Path Core Element

IPCE: International Path Element

NPE: National Path Element

CP: Customer Premises

The following categories are determined for the path lengths:

Path length, km	*L<500*	*500<L<1000*	*1000<L<1500*	*1500<L<2000*	*1500<L<2000*
I	*1*	*2*	*3*	*4*	*5*

Table 7.2.1

According to the length categories, the availability ratios and allowed outage parameters of the constituent sections of the entire link can be determined by the following equations and table:

$UR_{jS} = A_{jS} + i \cdot X_{jS}$ for standard category;

$UR_{jH} = A_{jH} + i \cdot X_{jH}$ for high quality category.

Path element	**Performance category**	**Mean value (·10^{-4})**		**Worst case (·10^{-4})**	
		A_j	**X_j**	**A_j**	**X_j**
IPCE	standard	0	15	40	35
	high	0	3	8	7
NPE	standard	0	20	52	47
	high	0	4	12	9
ICPE	standard	0	20	52	47
	high	0	4	12	9

Table 7.2.2

If the connection consists of *N* independent sections, the following equations can be applied:

Mean value: $U_{MN} = \sum_{n=1}^{N} u_{mn}$

Worst case value: $U_{WN} = U_{MN} + \sqrt{\sum_{n=1}^{N} (u_{wn} - u_{mn})^2}$

where u_{mn} and u_{wn} are the unavailability ratios of the individual sections, provided the sections unavailable ratios are according to the Standard Deviation.

7.2.4. Possibilities to increase the availability

Referring to the parameters affecting the availability described above, the possibilities to increase it are as follows:

Network planning phase

- Choosing the most advantageous network topology, minimising costs; Careful selection of network interconnections; Multiple access of network nodes via alternative routes; Application of dual-homing;
- Application of various self-healing architectures;
- During implementation planning, the allocation of equipment to secure places; Ensuring high reliability power supply and the specified environmental conditions;

Realisation phase

- application of properly reliable equipment that meet the availability objectives of the system

During operation

- Application and operation of properly planned network management systems for prompt fault detection and fast network reconfiguration;
- Increasing the efficiency of maintenance activities by applying properly skilled staff; Minimising the number of faults due to human mistakes;
- Modernisation of out-of-date, overused equipment; Reconstruction of old networks; Application of new technologies;
- Accurate planning of fault elimination processes;
- Application of spare parts stock with appropriate quantities; Optimal distribution of the spare parts.

7.2.5. Planning the number of operational staff

In large networks significant savings can be achieved, if the optimal number of repairing teams is determined by knowing the expected failure rate (λ), in such a way that the sum of losses due to failures and the maintenance cost of the teams is minimal. The optimisation can be done as follows:

If the network repair rate is μ, and the deviation of average repair times is exponential, the *W(N)* mean service outage time is as follows:

$$W(N) = \frac{\frac{\left(\frac{\lambda}{\mu}\right)^N \cdot \mu}{(N-1)! \cdot (N \cdot \mu - \lambda)^2}}{\sum_{k=0}^{N-1} \frac{1}{k!} \cdot \left(\frac{\lambda}{\mu}\right)^k + \frac{1}{N!} \cdot \left(\frac{\lambda}{\mu}\right)^N \cdot \frac{N}{N \cdot \mu - \lambda}}$$

The optimal number of teams (N_{opt}) is at the minimum of the sum of costs of fault repair and the income losses due to traffic outages, according to this formula:

$$N_{opt} = \min_{N} \{N \cdot C + \lambda \cdot W(N) \cdot F \cdot\},$$

where C is the annual cost of the maintenance of a team, and F is the income loss for a time unit.

The model can be refined by the priority orders of fault repairing of simultaneous faults, i.e. the higher priority traffic outages can be eliminated faster by interrupting the lower priority fault repairing activities.

It can be observed that using few repairing teams increases the loss due to traffic outages, while using too large repair staff causes unnecessary expenses on manpower. The size of operational equipment store can be optimised similarly.

The costs of repair and the maintenance staff can be reduced by several technical solutions, protections. In this case, the minimum of the sum of the investment cost of the network protection and the income loss due to the lack of protection has to be found. The technical solutions increasing the availability are described in Chapter 7.3.

Translated by Róbert A. Horváth

7.3. Network redundancy concept

(Editorial Chapter)

The structure of the network and the different self-healing solutions were discussed in chapter 4. In subchapter 4.1-4.4 traffic-routing methods where shown which can offer connectivity in the case of a single failure in any part or item of the system using mashed or doubble-ring networks. If there are at least four switching points in the networks than they can also transmit information from any source to the request sink in the case of more than one failures, but in this situation the quality will be impaired due to the higher congestion.

There are also an appealing solutions where the redundances are not only in the phisical layer, but in the higher layer too, supporting the information transfer in the case of the outage of any section. Here we can use the intelligence incorporated in the third and higher layers offering a routing strategy fitted to the existing situation. This was also mentioned in chapter four, including the planning of it.

Modeling the network with graphs it is possible to have a proper view on the possible failure tolerances of the network. Here (1.10)you can see that extracting any path you can find several other possibility to establish the requested connection. It has a furthur advantage namely mathematical model derived from the graphs gives a good background to calculate the optimal rout of traffic transportation. In the course of planning and realisation it is possible to prepare the routing strategy for every failure. So in some milliseconds the rerouted networks can achieve any terminal points. More details can find in 1.10 where the mathematical background was described by András Recski and Peter Laborczi.

Redundances can be planed and realised not only in the network structure but also in the unit or equipment level. In the case of equipment reservation we have two possibility. In the first case two equivalent units are working in parallel but the bit-stream is led on one of them. Only in the case of outage takes over the task the other one. There is no down-time so the user is not disturbed by the transition time. The maintenance staff will be informed about the failure immediately so they can change

the faulty unit. The engineering of this system is based on the availability requirements.

The goal of the maintenance organization is to minimize the down-time but in the same time the travelling time of the maintenance staff should be as low as possible. The number of the maintenance engineers, technicians is also a critical point because their salary is influencing the budget of the company. All this requirements can be fulfilled if the down-time of the network is much shorter than the fault-clearing time. This condition can be realised if they are in any position redundant units, transmission, pathes,power supplies which can be switched on or if we are using network-structures which can be automatically transmit the traffic in any other way for example selfhealing networks.

The availability is the ratio of the operating time (T_0) to the whole calendary time (T). It is characterising the available time for telecommunications. The abbreviation of this ratio is ($A=T_0/T$). *A* must be near to 1 in general 0,999 offers a good service. In engineering or calculation to handle a number extremly near to 1 makes difficulties. Therfore the use of down time ratio is preferable

DTR=1-*A*

This can be used easear especially if we take in account that roughly there are 10000 hours (exactly 8760) in a year. So 10^{-4} DTR means that there is less than one hour downtime. 8 hour downtime equalls $10x10^{-4}$, which is an acceptable limit. If the telecommunication is a part of a technological process,and has critical task or purpose in it than lower limit should be defined.

The availability is influenced by the relaibility and the fault-clearing time. For calculation purposes the relaibility can be characterized by the failure rate or with other words with the average (meanvalue) of the time between two consecutive failures. We are using μ for characterizing this average the dimension of it is time. In some cases the use of n can be preferable the average number of failures in a given time intervall. It’s dimension event/time. The fault clearing time is L. In the case of authomatic overswitching it is in some minutes or less. If the operation can be restored only by local repaire or change than the downtime is equall with the one way travelling time+ repaire time

$L=L_t+L_0$

Using this letters

DTR=nL

If n is given in event/year and L in hours/ event, then DTR=nLx$\frac{1}{8760}$,which is in general a small number. Therefore in the practice some times it is given in fit –5 which makes

DTR=$10^{9.}\frac{n.L}{8760}$ 10^5n.L (fits)

For calculation the number of the maintenance staff or maintenance groups we must use the double of the travelling time, because in general before coming the next task, they return to the basis. If the maintenance is organized based on real time management using mobilecommunication, than the tarvelling time can be reduced. In that case it is possible to send that group to the next failure which is in the nearest position to the required place and they can reach it by car quite rapidly. This system is extremly usefull if in the car can be find hardware and spare unit for any possible failure or at least for the most often occuring problems. There are computer-programmes which support the optimizing the amount of spare units and the routing of the cars. The telecommunication network is based on the use of highly reliable equipments where the propability of two failures on the same day in a maintenance area is less than 10^{-4} so the consecutive fault clearing is acceptable.

If there is no automatically switched redundancy or self-healing structure than the engineering is based on DTR=n(L_1 +L_0) overage workingtime should be minimum. From this can be calculated the optimal number of the maintenance group, the area belonging to one maintenance group minimizing the travelling time and giving a limit of three hour down-time. In general circles with 50-60 km radius makes possible an acceptable service. Depending of the amount of equipments can be defined the number of groups on this bases.

If there are new systems in the network on the territory the avalability must be repland. The new plan must be fited to the failure rate of the new systems and to the reduced fault-clearing time. If the network is overdimensioned than the failure is not causing disturbing traffic congestions so that the outage is not in close co-relation with the service impairment. In the acces network there is no alternative route and is

not possible overdimensioning so here a special fault monitoring can help in offering a service-level, defined by the SLA (Service Level Agreement).

In special cases (banking, technology supporting, ambulance, emergency services) the terminal can be connected to more than one switching center, so the failure of the acces line or the local switching or routing point can bypassed. An other method is to have simultaneously wire-line and wireless connection substituting each other.

Summarizing we can see that improvement of the network availability can be achieved by higher investments or maintenance costs. The customer can decide that which is minimum availability which is necessary for his job. Compairing it with the higher tariffs or rental price he can find the best compromise. Here it must be emphasised that cost (A) is not a linear function. (Figure 7.3.1)

7.4. Network Element Mangement and Network Management

József Wiener, author

Kornél Terplán, reviewer

In the last decade, more and more services are added to the traditional telephone services, and demand for transfer of high-bandwith data like video has increased The ability to deliver voice, video and data at higher speeds are becoming a critical requirement. Now, telecommunications networks have to integrate different computer networks with different bandwith demands. The growing Internet, its different nature as of the traffic also generates new reguirements for Network Operators and Service Providers.

In order to be able to carry different kinds of data at a faster rate, quite new technologies and significant improvements to the existing technologies and protocols have been and still are beeing introduced. Standards for Synchronous Digital Hierarchy (SDH), Asynchronous Transfer Mode (ATM), Digital Subscriber Line (xDSL) are well extablished, but implementations are at different stages at different Network Operators. Due to this facts and the rapidly changing demands, migration from the plain old telephone service (POTS) has to be realized very quickly.

Recently, fiber optics, wireless mobile communications have added their own complexity and speeds quite different from that of POTS. Telecommunication equipment shall be able to handle a variety of traffic in natures, speed instaed of only voice.

The complex and diverse equipment and service portfolio, their fast change makes controlling networks, resources and services a very complicated task. This problem can be solved by using modern computer technology and by as much automatization as possible.

7.4.1. What is Network Management?

Probably anyone who has worked with a telecommunications network has a different concept on this subject.

Telecommunications Network Operating Organizations have invested significant amount of time, money and human resources into building their complex telecom networks that need to be maintaned. Network Management can be identified as the process of controlling the telecom network in order to maximize its efficiency and productivity. This process includes data collection from the network (either manually or preferably automatically), processing the data and presenting it to the staff operating the network. As the importance of Service Management is increasing, Network Management shall also pre-process data for Service Management, and shall transfer it to higher levels (Service Management and Business Management) Systems.

During the last few years, there has been a major paradigm change in telecommunications. In many countries, deregulation led to strong, sometimes very agressive competition amongst telecommunication service providers. Deregulation also allowed service providers to expand their activities outside their traditional areas, even beyond their own national borders. This globalization tendency has also increased competition.

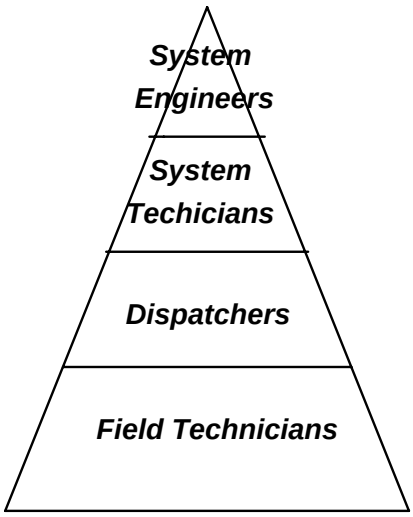


Figure 7.4.1. The Skill pyramid

Proprietary network management cannot provide interoperability between diverse technologies, equipment and networks deployed by different manufacturers and proprietary management solutions having different features and different level of management capabilities. The gap can theoreticially be overbridged by human resources. Hovewer, the speed that can be provided by human resources is often not fast enough to provide the contracted level of service for the customer.

As it is known, there exists a so called "skill pyramid" (Figure7.4.1). This means that the higher skill is needed, the less staff is available. Also, the expenses of

training and salaries are increasing for the higher skill. The diverse and more and more complex technologies would need more skilled staff, but economic operations and efficiency would require less - no doubt that the requirements for more economic operations are the stronger ones, and companies can hire less skilled persons. Consequently, less and less staff shall run the bigger and bigger and more and more complex networks as well as services.

As a consequence, Network Management (NM) is becoming more and more critical. The essential idea behind NM is to replace human resources by computers in order to automate and speed up processes.

Proprietary network mangement solutions are usually working well with the manufacturers own equipment, but cannot provide interoperability between diverse tecnologies, diverse network management and service management solutions. This is driving the need for standard NM solutions important and urgent.

Legacy (proprietary) network management systems generate another problem, as high amount of money and human resources have been invested into these systems. It is difficult to discard solutions that are working and into which heavy investments were made.

This necessitates that legacy systems will coexist with the new standard network management solutions, and standardization work shall take into consideration this necessity.

7.4.2. Process Orientation

The Importance of Service

The core of a communications service provider is the service itself. The key objectives are ‘more for less’ - faster service introduction and provisioning, improved quality of service at a lower cost. These objectives can only be achieved through automation of customer care and operational support processes, and a strong automated linkage between the management of customer service offerings and the underlying networking assets. The level of customer service provided and the level of automation in the current environment of almost all Service Providers is much lower

than expected and lower than what providers need to remain competitive. Business process-driven approach

.Service Providers and Network Operators need to automate their business processes, which means information needs to flow from end-to-end across many different systems. All the activities, all the processes inside the Network Operator and Service Provider shall support the Business Processes; Network Management, Service Management, Marketing, Procurement, etc. shall collaborate with these Business Processess

A Business Reference Model

The Business Reference Model shown in Figure 7.4.2 is the basis for management. It illustrates the principal points of contact between a service provider, its customers, its suppliers and other service providers. A wide range of automation and integration opportunities exists among the business roles and relationships shown.

7.4.3. TMN (Telecommunication Management Network)

In telecommunications, and especially in the telecommunication industry, TMN (Telecommuncations Management Network) is a loosely used term covering all kinds of network management solutions. In the strick meaning, hovewer, TMN refers "only" to network management solutions that conform to the standards of ITU-T.

The legacy management solutions of telecommunication equipment manufacturers were good enough for limited services and limited geographic are covered by their products. Because of the regulatory and protected environment, network operators & service providers had time and could control the introduction and implementation of new techologies and solutions.

The TMN model is a way to think logically about how the business of a service provider is managed. The TMN model consists of five layers, usually arranged in a triangle or pyramid (Figure 7.4.3). Business Management is at the top, Service Management is the second layer, Network Management is the third layer, and Element Management the fourth layer with the physical network elements is represented in the bottom layer.

The concept is that management decisions at each layer are different but interrelated. For example, detailed information is needed to keep a switch operating (at the element management layer), but only a subset of that information is needed to keep the network operating (e.g. is the switch operating at full capacity). Working from the top down, each layer imposes requirements on the layer below. Working from the bottom up, each layer provides capabilities to the layer above.

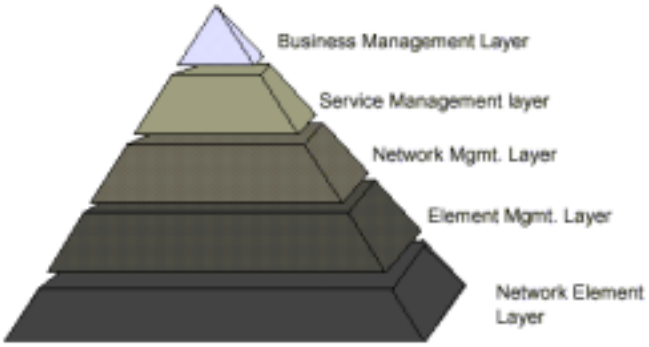


Figure 7.4.3. The TMN Model

TMN specifies a layered architecture for management of telecommunications networks. It deals with the monitoring, control and coordination of the resources in the network (resources are any components of the network, providing services - including equipment, software, hardware and customers). Some functions of TMN are:

- Remote management of systems components, hardware and software involved in providing services to customers - like voice, video, IN services, Internet, data communications, etc.
- Providing easy interfaces and easy interactions with customers in order to configure their services. This interface has to take into account the different skill levels of the end users as well as the operators.
- Providing automation in discovering and fixing problems. This also includes self-healing and self-correction in networks.
- Achieving seamless integration and management of legacy equipment and protocols with the new equipment and protocols.

Not only hardware, but also software is used to enhance the functionality of telecom equipment. This software and TMN (software) applications themselves have also to be managed.

7.4.4. Systems Management Functional Areas

To perform the managament tasks, network management consists of five functional areas:

- **F**ault management covers the detection, isolation by diagnosis and analysis, and correction of problems in the network. Fault Management also covers reporting and problem tracking e.g. by Trouble Tickets.
- **C**onfiguration Management is used to keep track of resources in the network. This includes not only configurating equipment, but also covers areas such as view management, topology management, software management, inventory management and provisioning.
- **A**ccounting Management covers the usage of resources, controlled collection of data on the usage and charging for tha usage.
- **P**erformance Management covers perfomance data collection, analysis of performance data and reporting of problems. Performance management is also concerned with the behaviour and evaluation of the effectiveness of resources.
- **S**ecurity Managament function covers detecting and reporting security violations, creating, deleting and maintaining security-related services such as encryption, key management and access control. Distributing passwords and secret keys is also a function of security management.

These five areas are sometimes referred to as FCAPS, and the five categories are called Management Functional Areas (MFAs).

7.4.5. TMN Architecture and Functional Grouping

TMN needs an application architecture for the software of network element, network and and service management software. Management functions are generally performed by distributed computing.

For the purpose of realisation, the ITU-T Telecommunication Network Management (TMN) information architecture shall be decomposed to a set of traceble requirements and explained in the form of possible to a way of considering practical implementation issues. This division makes possible the use of the standards in the industry as a set of requirements for componentized distributed computing in general.

Due to the endorsement of a layered TMN architecture in the industry, a tele-communications management building block

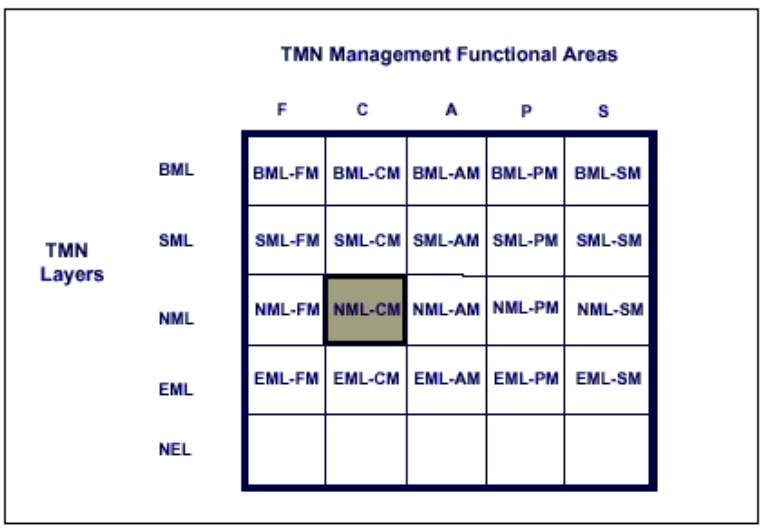


Figure 7.4.4. Layering and Grouping

must contain functions from only a single TMN layer and from a single TMN management functional area. This granularity requirement, constraining the maximum range of functionality of a telecommunications management building block, is here stated in terms of well-accepted layered architecture from the telecommunications management industry.

TMN formally requires a separation between the layers. This layers are separating the main areas of the business as operating the network (Network Elements & Element Management), managing the network (Network Management), Managing the Services provided for the customers (Service Management) and Managing the Business (Business Management). This well known layering is depicted on Figure 7.4.4.

In actual practice, it will often be found that each MFA would be implemented most conveniently as several building block types. For example, EML-FM contains element managers both for network alarm surveillance and for network test equipment. These element managers have different clients on the NML and different scaling and distribution issues. Implementing them separately makes it possible to deal with each of them in the most optimal manner.

7.4.6. TOM - Telemanagement Forum's Telecommunication Operation Map

The TMN model is simple, although its implementation is complex. The big number of standards now available that address the various interfaces between management systems sometimes makes it difficult to see and appreciate the whole picture. TMN Management Functions provides a structure and decomposition of functions for all of the layers. However, those ITU-T standards that specify information models and interfaces have been mainly concentrated on the management and connection of resources to the Network Element Layer. Until recently, little attention has been given to interface specifications and information models within the TMN. Consequently, it is difficult to apply the standards to a complete business case, such as for the procurement of a specific Operations Support System. It is also difficult to apply a customer centric focus on the processes that really respond to the customer needs.

The Telecom Operations Map (TOM), using the TMN model as a foundation, addresses operations support and management for any communications service from a top down, end-to-end process and customer oriented standpoint.

The TOM (Figure 7.4.5.) serves as the blueprint for process direction. It is also the starting point for development and integration of Business and Operations Support Systems (OSS). It consists of:

- A high-level view of Communications Operations processes, sub-processes and activities that is top down, customer-centric, and end-to-end focused;
- A high-level identification of the primary end-to-end processes of fulfillment, assurance, and billing, and sub-processes;
- Illustrative examples of process flows that show end-to-end processes;
- A more detailed view of the functions of each sub-process, including activities of each sub-process box, as well as its inputs and outputs to other sub-process boxes.

The Telecom Operations Map uses the layers of the ITU-T TMN model to organize core business processes, but divides the Service Management layer into 2 parts:

The first is Customer Care, the second is Service Development and Operations Processes. In the simplest sense, this division reflects differences between processes triggered by individual customer needs from those applied to a group of customers subscribed to a single service or service family. These processes are responsible for ensuring that the network and

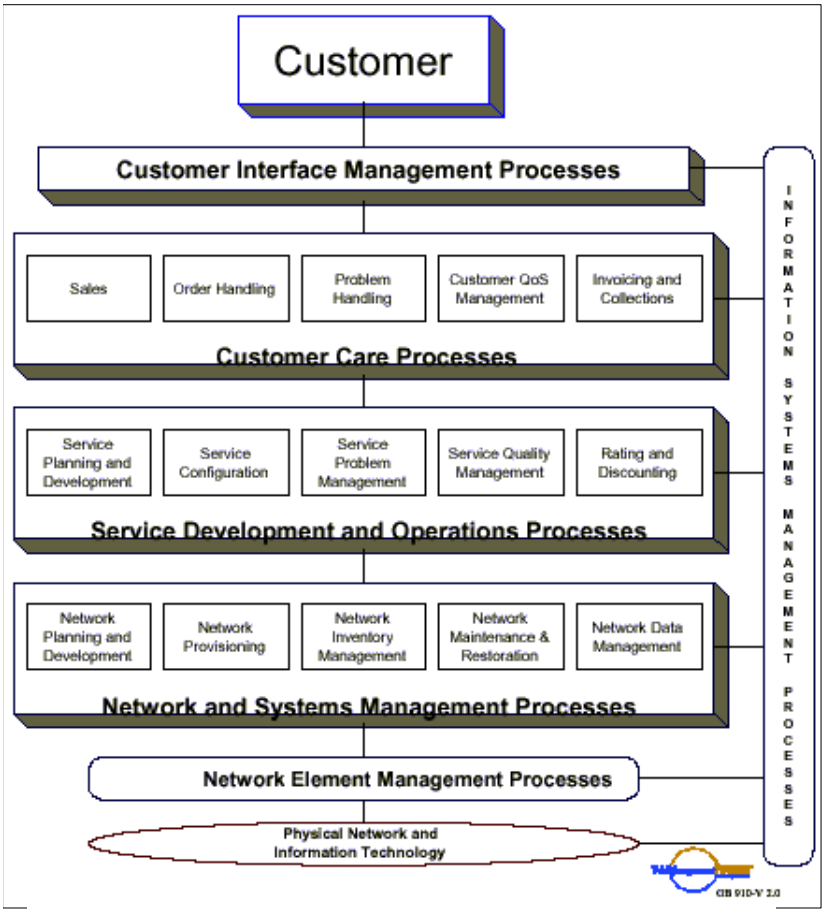


Figure 7.4.5. Illustration of TOM's Processes

information technologies infrastructure supports the end-to-end delivery of the required services.

Network and Systems Management is also the integration layer between the Element Management Layer and the Service Management Layer. Its basic function is to assemble information from the Element Management systems, and then integrate, correlate, and in many cases, summarize that data to pass on the relevant information to Service Management systems or to take action in the network.

7.4.7. Network Management Operations Map

As it was shown,Network Management Processes are part of the TOM. Network Management Processes are detailed further in the Network Management Operation Map.These processes and the relationship to TMN is shown on Figure

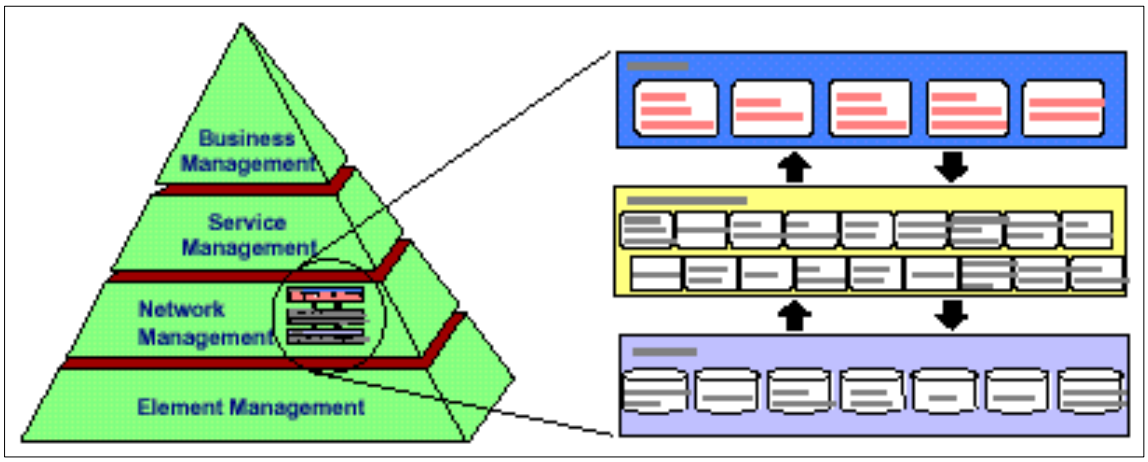


Figure 7.4.6.
Positioning the Network Management Detailed Operations Map within TMN

7.4.6

7.4.8. Element Management and Network Management

Element Management and Element management systems (EMS)

The telecom network consists of Network Elements (Equipment). *Equipment* or *Network Elements* (NEs) is the short term for the basic infrastructure, i.e. the hardware and software components of the network. In the usual case, equipment is vendor-specific to a large extent. For billing, this means that the raw accounting data

is delivered often in a vendor-specific format and reflects the functionality of the elements actually used.

The *Element Management System (EMS)* represents the hardware and software components used by the Service Provider or Network Operator to manage one or more Network Elements (NEs). The EMS provides management across a subnetwork or a single NE, typically across a single vendor equipment or collection of single vendor equipment. The NMS performs management functions across the Element Management Layer (EML) of the TMN. Some examples of these management functions include provisioning of NE resources and collection of NE faults.

Full separation of the MFAs on the EML is problematic with the current generation of network elements (NEs). Some NEs can connect to only one or two management facilities. This will require EML building blocks managing different aspects of an NE to share an element of common infrastructure. This common element will communicate directly with the NE then forward the information appropriately to the various EML building blocks according to their respective management interests. Additionally, the EML mediation function, which brings various NE management protocols from different vendors to a common view, might be performed most conveniently at a single point. This common single point of NE contact can be implemented either as an EML mediation building block, which will require open contracts, or as infrastructure, which can be vendor proprietary.

Network Management System (NMS)

The *Network Management System* represents the hardware and software components used by the Service Provider or Network Provider to manage their networks as a whole. The NMS provides an end-to-end network view of the entire network enabling management of the NEs contained in the network. These NEs managed across the network are typically provided by multiple vendors. The NMS performs management functions across the Network Management Layer (NML) of the TMN. Some examples of these management functions include connection management and circuit fault correlation.

Network Management is more than just a mediator between the EML and SML. Network Management processes have their own responsibilities; for example,

- Network Planning and Development (assuring complete infrastructure exists);
- Network Provisioning (implementing the infrastructure);
- Network Inventory, maintaning status of network assets;
- Management (implementation and administration of the physical network);
- Network Maintenance and Restoration (assuring availability and maintenance of the infrastructure) and
- Network Data Management (collects data to manage the network and provide billing records).

The important issue is that management responsibility will be placed at a level where adequate information is present, instead of shifting all responsibilities to Service Management.

The Network and Systems Management processes manage the complete service provider network and sub-network architecture.

NML-EML Interface

The composition of today's networks has contributed to the complexity of managing these networks. These networks are commonly composed of network elements provided by various vendors. The task of interoperability extends beyond the network element layer up to the management layers, to include interoperability between multi-vendor Element and Network Management Systems.

Service Providers have also taken advantage of technological advancements in transport network equipment. It is not uncommon for service providers to deploy next generation, multi-technology network elements, (or "hyrbrid NEs"), as these network elements provide new services and optimal network resource utilization. However, existing network management solutions that have been specified to date apply only to a specific technology.

There is an industry demand for a full-featured, commercially available, scalable and non-proprietary network management solution, where multi-vendor, multi-technology management systems

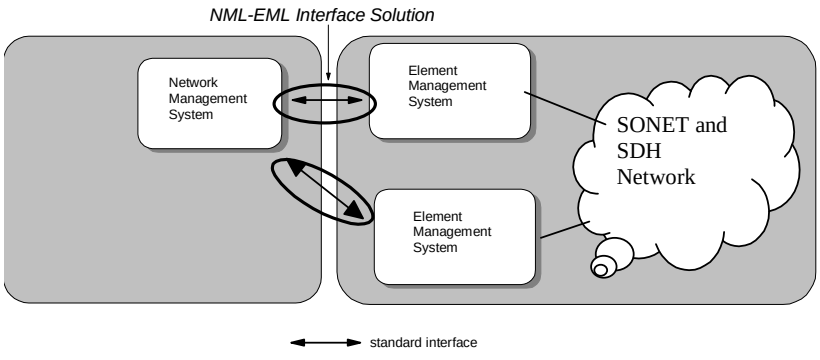


Figure 7.4.7. NML-EML Interface

interoperate in an open architecture environment.

The *NML-EML Interface (Figure 7.4.7.)* represents the communication data and exchange mechanism between the management system(s) that deploy the NML and EML functions of the TMN. A Network Management System (NMS) that performs NML functionalities may communicate with one or more Element Management Systems (EMSs) that performs EML functionalities via the NML-EML Interface.

Structuring the EML and NML Layers

Traditionally, Element Management level within the TMN Architecture has been used as the boundary for containing the technology specific management capabilities whilst giving a generic view of the network to the higher management layers (e.g., NML, SML etc.). Such a boundary is adequate for managing individual technologies but leaves much of the integration across technologies to be undertaken by the Service Providers/Network Operators. Increasingly, Service Providers/Network Operators seek solutions from the Equipment Suppliers in the form of complete Sub-Networks. Such Sub-Networks may consist of a range of technologies (Routers, Switches, etc. in an IP network) from multiple vendors.

In practice, the distribution of functionality may vary significantly between Element Management and Network Management Layers. One example is shown on

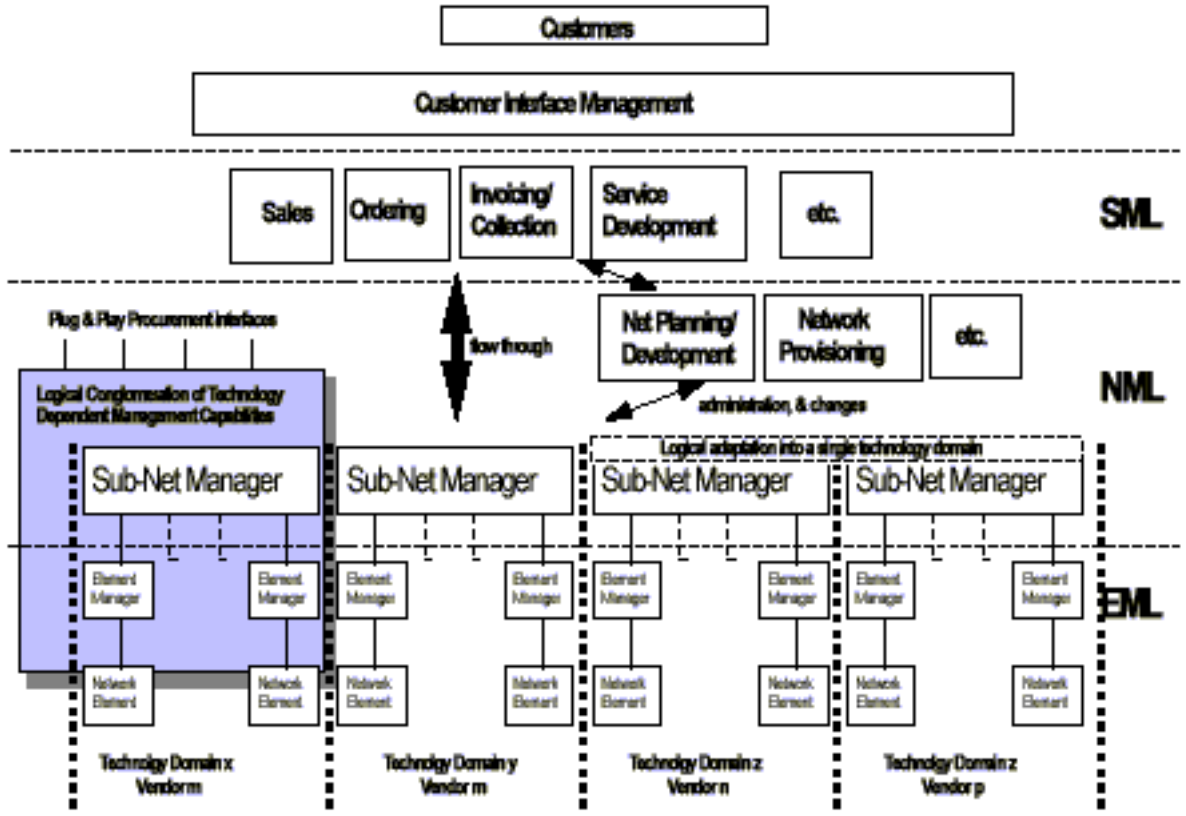


Figure 7.4.8. Structuring the Network Management Layer

Figure 7.4.8. The functional areas can be encapsulated as blocks of functionality within the TOM. The chosen functional blocks can reflect the distinction between generic and technology-specific management indicated above. The functional blocks shown in Figure 7.4.8. have been chosen to distinguish Network Management capabilities associated with Element Managers (at the node level) and Sub-Network Manager(s) (for some managed network area or domain).

The partitioning Network Management Layer into Generic and Sub-Network specific components and providing ‘ plug and play’ type interfaces at the Sub-Network level can also be reached this way.

Note that this is only one possible structuring. Currently, there is no industry wide agreement on such a partitioning. Procurement by Service Providers of managed network technology is often based on a combination of Element Management and some aspects of Network Management packaged into this type of Sub-Network Management. This allows the managed Sub-Network domain to be accessed as a network area, rather than just a series of individual network nodes. The Sub-Network might be defined:

- on the basis that it employs a particular network technology (e.g. SDH or ATM), with its associated management, and is procured as a package;
- on the basis of geographical partitioning;

This can be done as specific areas of management functionality are supported, and different organizations have different needs.

7.4.9. Network Management Relationships

The following interactions impact Network Management processes and directly drive the need for interface specifications in the form of information agreements that may need to be automated.

Interactions with Service Management

This is one of the primary relationships for Network Management and acts as the main source of requests for information and actions to execute tasks. Service Management is responsible for managing the customers’ perspective for each individual service provided, normally against some type of contractual agreement.

Thus, its purpose is to ‘ act on behalf of the customer’ for interactions with Network Management.

Interactions with Suppliers or with Supplier-provided Equipment

Most traditional Service Providers own and operate networks in order to deliver their services. Certainly, the service delivery chain will always include at least one Provider which takes on this Network Operator role. For these Provider/Operators, the network operations task is an internal business function rather than a point of external interface. However, since most Service Providers do not manufacture their own network equipment, they are reliant on the equipment suppliers, from whom they procure, to help them achieve their automation goals. The ability for devices to be configured in a common way, for example, or to provide alarm or performance data using common formats and terms, is critical to achieving the full benefits of service and network management automation. To get the most from automation efforts, procured equipment must be able to receive and act on common high-level instructions, and deliver performance and usage-related information in a common way, that meets the Providers' requirements.

Interaction with Customers

Most Service Providers see a need for automated management links with their Customers, at least with some types of Customers, and/or for some types of services. These interactions occur mainly with Service Management, which act as a proxy for the customers’ needs to Network Management.

Interactions with Other Providers

World-wide alliances and regulatory actions are generating increasing volumes of interactions between Service Providers. Today, these often involve manual intervention, representing an unacceptable cost and often significantly degrading service quality to the Customer.

Some of the interactions between Providers may be similar in content to the interactions between a Provider and a Customer. However, it is likely that the volumes of transactions, the level of detail required, and the speed with which

information needs to be exchanged between Providers will dictate substantially different implementation agreements.

7.5. Service Management In Telecom Networks

József Wiener, author

Kornél Terplán, reviewer

As the communications industry continues to evolve with deregulation and liberalization, service providers are under increased pressure to deliver a broadened set of services at a competitive price. As a result, service providers must deliver these services in an cost-effective and timely manner. The ability to effectively manage these networks become key in retaining existing, as well as acquiring new market share. However, the complexities surrounding today's networks present challenges in achieving the effective network management goal.

7.5.1. Role of Service Management

The core of a communications services provider is service. The key objectives are 'more for less' - faster service introduction, improved quality of service at a lower cost. These objectives can only be achieved through automation of customer care and operational support processes, and a strong automated linkage between the management of customer service offerings and the underlying networking assets.

Both the level of customer service provided and the level of automation in the current environment of almost all Service Providers is much lower than expected and lower than what providers need to remain competitive. Many Service Providers are now actively engaged in re-engineering their business processes for maximum efficiency and effectiveness. The effective exploitation of this network infrastructure, whether directly operated or outsourced, is an integral part of the service delivery chain and directly influences the service quality and cost perceived by the end customer.

The idea is that the End Customer develops telecommunication service quality requirements necessary to operate their business. These requirements are brought to the Service Provider and the two parties begin to assemble the optimum set of SLA parameters and values for the services.

The two parties (Customer and Service Provider) may be an ’end’ Customer and their Service Provider (SP) or two Service Providers, where one Service Provider has the Customer role buying support services from another Service Provider (e.g. who may be acting as a network operator).

The agreed-upon SLA requirements flow down through the organizations of the associated SP(s) and become the basis for internal management processes and QoS values. Customer satisfaction is improved by identifying the implications for supporting the service by the internal support infrastructure(s) of both the SP and the Customer.

7.5.2. Quality of Service (QoS)

By definition, Quality of Service (QoS) is the collective effect of service performances which determines the degree of satisfaction of a user of the service. The quality of service is characterized by the combined aspects of service support performance, service operability performance, service integrity and other factors specific to each service (ITU-T Rec. E.800).

The SLA QoS parameters support a contract between two parties. It is important to note that there is a distinct difference between the user/service QoS requirements defined in the SLA and the network level QoS.

Quality of Service (from SLA point of view) is the measure of the service quality defined for a service and provided to a customer. QoS is the definition of the performance parameters used to assess service quality. The parameters are usually associated with a specific service or service type. Traditionally, the term QoS is used to refer to performance related parameters. Some use QoS to mean the quality of service for all aspects of the service, e.g., network performance measures and Completion On Time, Call Pick-up Time, etc. QoS can be subjective, e.g., is a call easy to hear for voice, or objective, e.g., Cell Error Ratio for ATM.

Defining QoS is easiest with digital circuits. QoS for IP Services is getting a lot of attention, since it is a connectionless service that is hard to measure and since QoS for IP Services came from the IT arena, i.e. the “best effort” delivery model.

The SLA should include clear and unambiguous definitions of the following:

- The measurable QoS metrics and parameters that can be guaranteed by the SP for a specific service in terms that the Customer can understand and agree to.
- Service performance measurement method, measurement period, reporting period and reporting frequency.
- Customer and SP responsibilities, e.g. maintaining relevant hardware and software.
- SP procedures to be invoked on violation of SLA guarantees.
- Any conditions that affect operability/commitment to support.
- Selection of the type of reports associated with the service, specifying each report's contents, format, destination, conditions and delivery media.
- Service definitions for each service covered by the SLA.
- Process for handling the defined boundary conditions.
- Service cover time, i.e. the limits of SP support for different times of the day/week/month/year, etc.
- Dispute resolution procedures.

For any service the Customer should be able to select:

- Parameters to guarantee.
- Value ranges for the parameters.

7.5.3. Service Level Agreement

The SLA is a formal negotiated agreement (contract) between the two parties, and it is designed to create a common understanding about service quality, priorities, responsibilities, etc. SLAs can cover many aspects of the relationship between the Customer and the SP, such as performance of services, customer care, billing, service provisioning, etc.

Although an SLA can cover such aspects, agreement on the level of service is the primary purpose of a SLA. The focus is therefore on the management of the SLA and the Quality of Service (QoS) that is agreed upon in the SLA.

The parameter categories in the SLA are 1) technology-specific, 2) service-specific and 3) technology/service-independent. The Customer has two interests: 1) impact on the single user and 2) aggregate performance for a defined period.

An SLA, in many cases, is part of or an addendum to the contract with the customer. It defines the service provided and the set of metrics to be used to

measure the level of service committed against the level of service provided. Such service levels might include network performance metrics, installation completion on time metrics and intervals for new orders, availability, call pick up times at a work center, maximum periods of outage, average and minimum throughput, etc. The SLA also frequently defines trouble reporting and escalation procedures, reporting requirements and the general responsibilities of both parties.

7.5.4. Relationship to the Telemanagement Forum's Telecom Operations Map

Service providers must apply a customer-oriented and service management approach, using business process management methodologies, to cost effectively manage their businesses and deliver the service and quality customers require. To manage within a service provider's value chain, a common process framework is required. The TOM (Figure 7.5.1) uses the layers of the ITU-T TMN model to organize core business processes, but divides the Service Management layer into 2 parts: Customer Care and Service Development and Operations Processes. In the simplest sense, the division reflects differences between processes triggered by individual customer needs from those applied to a group of customers subscribed to a single service or service family. It also reflects the accountability for direct customer contact handling in Customer Care Processes and the critical need to focus on integration and automation of and in support of Customer Care Processes.

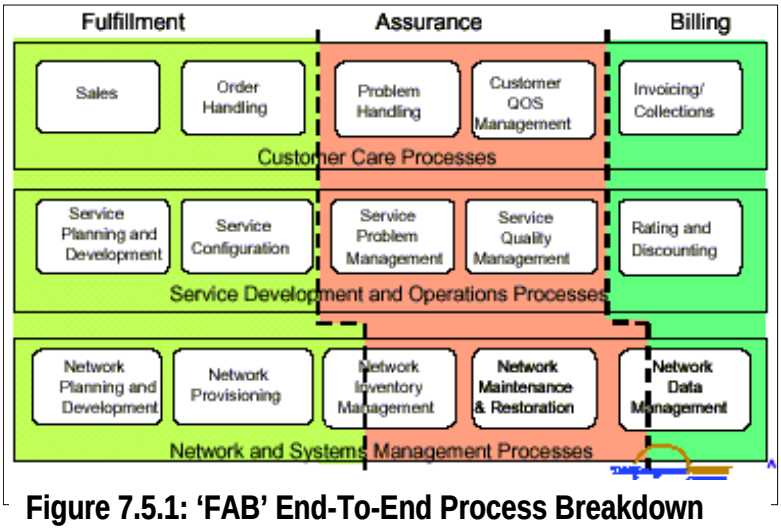


Figure 7.5.1: 'FAB' End-To-End Process Breakdown
(**FAB =** Fulfillment, Assurance, Billing)

Note: The interface to element management systems and the physical network are not shown for simplicity.

7.5.5. Service Development and Operations Processes

The Service Development and Operations Processes are on middle row of the TOM (See Figure 7.5.1).

These processes focus is on service delivery and management. Some of these functions are done on a one-time basis, like designing and developing a new service or feature. Other functions involve service capacity planning, the application of a service design to specific customers or managing service improvement initiatives, and are closely connected with the day-to-day customer experience.

Service Planning and Development Process

The process begins with the need for new service, feature, other concept/requirement, or a shortage of capacity. Triggers may come from customers, or from different departments of the company. It ends with introduction of the new service, feature, added capacity, including being able to sell, order, maintain, bill, report on and meet or exceed service quality, performance and cost targets.

This process encompasses:

- Designing capabilities to meet specified market need(s) at desired cost. This can be a new service, new feature, service enhancement, upgrade or maintenance related.
- Ensuring that the service (product) can be properly installed, monitored, controlled, and billed.
- Initiating appropriate process and methods modifications, as well as changes to levels of operations personnel and assuring required training is performed.
- Initiating any modifications to the underlying network or information systems to support the service requirements.
- Performing preservice testing to confirm that the technical capability works and that the operational support processes and systems function properly.
- Ensuering that sufficient capacity is available to meet sale forecasts.

Service Configuration Process

This process encompasses the installation and/or configuration of service for specific customers, including the installation/configuration of customer premises equipment. It also supports the re-configuration of service (either due to customer demand or problem resolution) after the initial service installation. Also,

reconfigurating the network to meet new demands and increase capacity may be part of this process, and such Service Configuration work would not be tied to configuring a specific customer instance for those cases. The aim is to correctly provide service configuration or re-configuration, including connection management activities, within the timeframe required.

Offering IP-based services, additional functions must be considered. Firewall, application services such as e-mail, web-hosting and their handling is important. Also, setting QoS and SLA parameters shall be performed.

Service Problem Management Process

This process encompasses reporting on service problems and trouble performance, isolating the root cause of service-affecting and non-service-affecting failures and acting to resolve them. Typically, failures reported to this process affect multiple customers. Actions may include immediate reconfiguration or other corrective action. Longer-term modifications to the service design or to the network or information technology components associated with the service may also be required. The aim is to understand the causes impacting service performance and to implement immediate fixes or identify quality improvement efforts required. The task include

- Isolate and resolve service problems
- Identify chronic failures
- Provide performance data
- Recommend service redesign, if appropriate
- Initiate escalation procedures
- Analyse service quality
- Generate reports about services

The process ends with the service problem rectified, improvements made, related development recommended or a decision made not to take action, and production of a root cause or other analysis reports.

Rating and Discounting Process

For a usage billed service, one of the essential activities of Rating and Discounting is to match usage to a customer record. As with other processes, some

providers provide rating and discounting functions for other providers as a service. For joint service arrangements, billing, invoicing, settlements and reconciliation between service providers may be involved.

The aim is to correctly apply charges, rate usage and to correctly apply discounts, promotions and credits. The process starts with registering a specific customer's identifiers for matching to usage and appropriate discounts, charges and/or credits. It ends with providing correct information for the billing invoice

7.5.6. Service Quality Management Process

This is a sub-process of the Service Development and Operations Processes, but due to its importance, it is dicsussed in this separate chapter.

This process supports monitoring service or product quality and cost on a service class basis in order to determine whether

- Service levels are being met consistently
- There are any problems with the service or product or any improvements are needed,
- The sale and use of the service is tracking to forecasts

This process also encompasses taking appropriate actions to keep service levels within agreed targets for each service class and to either keep ahead of demand or alert the Sales Process to slow sales. If improvements are required to the service or the infrastructure to maintain or improve service results, this process provides recommendations and tracks that approved developments are completed and/or that other required actions are completed. The aim is to provide effective service specific monitoring, to provide meaningful and timely performance information and to ensure service performance meets or exceeds commitments. This information can be used for specific customers (to internal management and customers, through the Customer QoS Process). The aim includes the monitoring, analysis, and reporting of service levels to meet SLA commitments or to meet standard commitments for the specific service or service class.

The Service Quality Management Process manages the service from first service to retirement of the service. Therefore, the process begins with Service

Introduction and includes the effective and efficient management and reporting of service results to meet or exceed the committed operations objectives.

The process include several sub-processes. Division of the process into sub-processes may depend on the Service/Network Proveider. Two of them (Lyfe-cicle Management and Maintaning SLA) are discussed in the following chapters.

7.5.7. The Life Cycle of the Service

To clarify the roles of the Customer and the SP, a Service and its SLA is divided into five Life Cycle stages: product/service development, negotiation and sales, implementation, execution, assessment. Each life cycle stage addresses specific operations processes in the *Telecom Operations Map* (Figure 7.5.1). The SLA Life Cycle provides a complete process description by delineating interactions between well-defined stages.

The Service Management processes form a longer periodicity lifecycle driven by the introduction, modification, and withdrawal of different service products (or ‘ classes’ of service). This lifecycle involves creating the specific policies, rules, process, and data templates used to configure and select service products the Customer Care process can utilize.

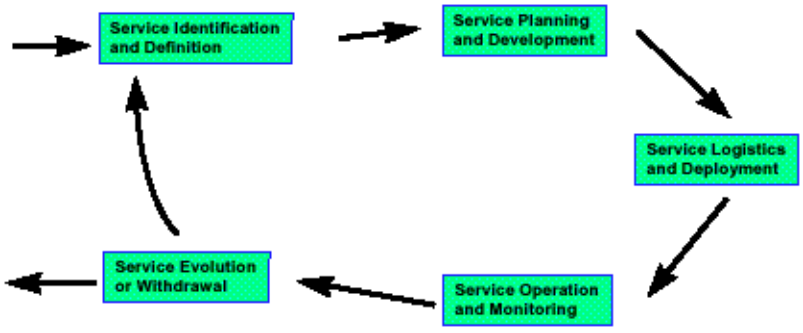


Figure 7.5.2
Typical Service Management Lifecycle

While there can be many combinations in how a particular company will segment and name their particular processes and methods, the overall lifecycle will generally contain many of the same steps. Figure 7.5.2. gives a typical view.

7.5.8. Maintain Service Level Agreements

Customers are interested first of all service-dependent metrics that are technology independent. This means, that technology-dependent metrics need not to

be included in the SLA, only if the service can exclusively be provided by one technology.

Service/technology-independent QoS parameters are those which are often (if not always) specified in a SLA. Examples include Percentage Availability, MTBF, MTTR, time to first yield, average call response time, etc. These are sometimes referred to as “operational performance criteria” and some are reportable by SPs to regulatory authorities on a regular basis, e.g. time to first yield.

Examples of service-specific QoS parameters

Some examples for service-specific QoS parameters are:

- **Voice telephony:** call connectivity and quality measures ABR/ASR/CCR/CSR/NER; network connection failures, Customer Affecting Incidents (CAIs), PSTN speech and 3.1 kHz audio loudness (speech level), attenuation, noise, crosstalk, echo, distortion; ISDN call set-up failures and delay, propagation delay (including differential delay between B-channels), G.821 error performance, premature release, release failure and delay, CLI reliability. With the increasing use of digital network technology, echo has become increasingly important, even for quite close destinations from a caller.
- **Voice over IP (VoIP):** delay and echo are two of the major hurdles to overcome.
- **Data:** BER, % EFS, errored PDUs, lost PDUs, UAS, Availability digital data parameters; loss, attenuation, group delay distortion, noise, impulse noise analog data parameters.
- **Facsimile:** image quality, character error rate, call cut-off, modem speed reduction, transaction time and availability.
- **Mobile telephony:** call completion rate, call dropout rate, noise, echo, distortion and availability.
- **Sound programme:** noise, crosstalk, stereo channel interference, distortion and availability.
- **Frame Relay Service:** It my include several parameters for access, plus per PVC parameters. For switched circuits, Customers negotiate other behavior based on the establishment of the SVC, Service Availability, Access Line Speed, Per VC Cell Rate (can be quite complex for non-real-time and real-time streams), Per VC Cell Propagation Delay, Per VC Cell Delay Variation, Per SVC Call Setup time, Per SVC Blocked Calls, Cell Error Ratio, Cell Loss Ratio
- IP-VPN Service - The following parameters may be part of the contract: Service Availability, IP Packet Transfer Delay, IP Packet Delay Variation (jitter), IP Packet Loss Ratio, IP Packet Error Ratio, Utilization.

Service Avilability

Service Availibility (SA) as a percentage (SA%) indicates the time during the contracted service at the Service Access Point (SAP) is operational. Operational means that thecustomer has the ability to use the service as specified in the SLA.

An event affecting the service at the SAP can be defined as an outage. The duration of the outage is outage interval. This concept is used for the unavaibility percantage (UA%) and service availability percentage (SA%) calculations as follows (in the simplest case):

- SA%=100%-UA%,
- UA%=(sum of outage intervals/activity time) x 100%.

  Service Availability has three dimensions:

- Time dimension
- Site dimension
- Functional dimension

Time Dimension of Service Availibilty

Service availability is not only a simple sum of all the availibilty durations. In some cases, some time periods can be excluded from the calculations. An example is, if a shop is closed during the night, and no communication is needed at all, this duration can be excluded from the SLA calculation. An other example is, if the Service Provider and the Customer agree in a Maintanence Window, when the

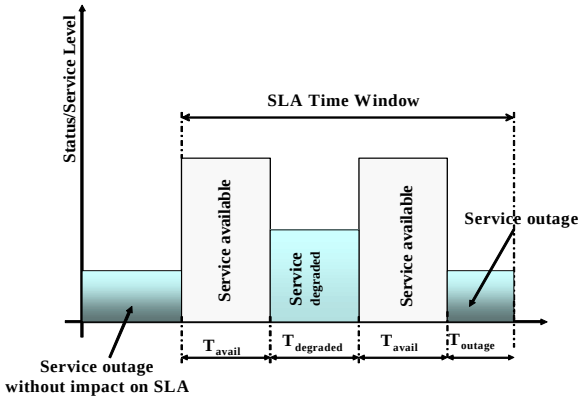


Figure 7.5.3. Time Division of Service Availability

Service Provider can perform scheduled maintenance work (conditions of having such Maintanance window shall also part of the SLA).

Figure 7.5.3. shows an example for this cases. There is also a situation on the figure, when the service is degraded for some reasons (e.g partial capacity or speed is available, performance degraded, etc.). Handling this situation in the SLA needs special attention from both parties. One possible solution is, eg. if this cases have some weights in the SLA calculations.

Site Dimension of Service Availability

Service Availibilty requirements may vary from site from site at a customer, depending on its importance, open/closed tiem, etc. This is shown on Figure 7.5.4.

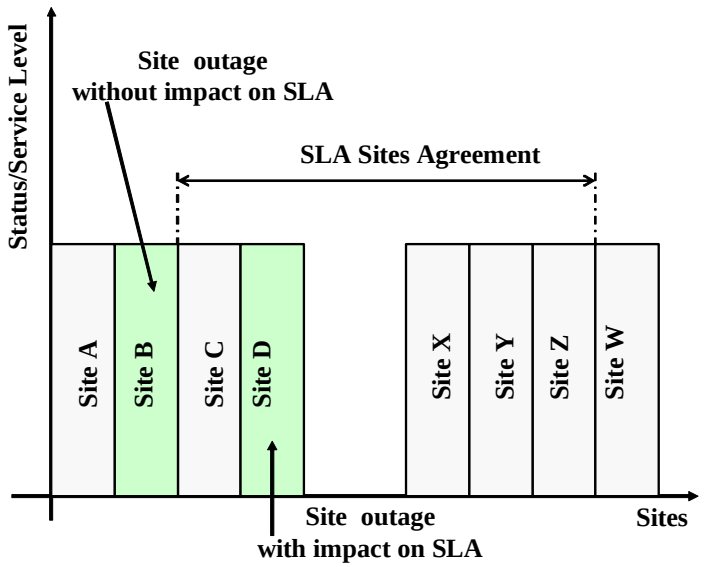


Figure 7.5.4. Site Dimension of Service Availability

Functional Dimension of Service Availability

Contracted and provided Service Availability may depend on the Function (Service) provided for the Customer. Some functions (Services) may be exluded from tha SLA, others may have different weights when calculating tha SLA (Figure 7.5.5)

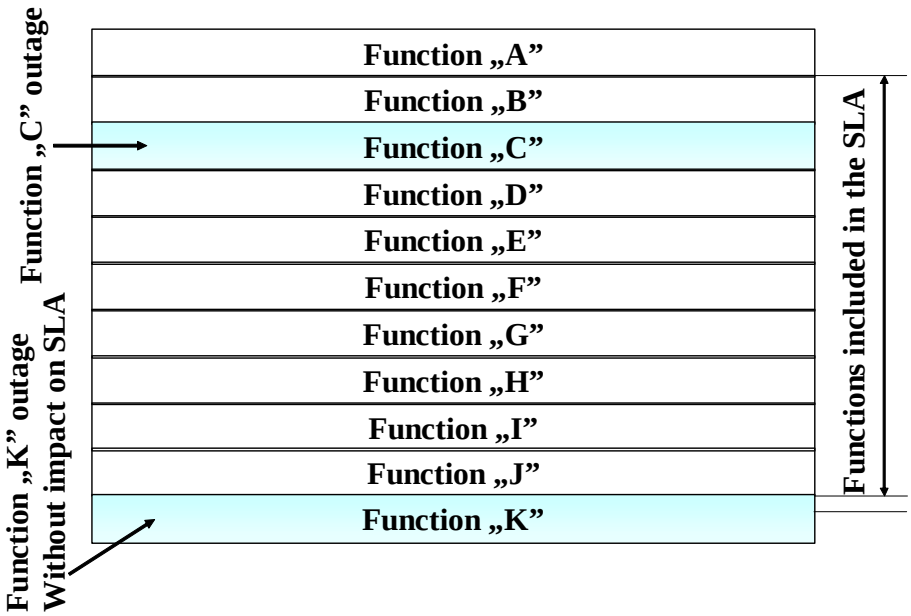


Figure 7.5.5. Functional Dimension of Service Availability

7.5.9. SLA Issues with IP-based Services

QoS is a key factor for the success of IP-based products and services. Hovewer, the approach is different from that of the traditional telecom services. This is due to the fact that it came from the IT world, and also as it is based (basically) on packet-switching.

Best Effort Service

This is basic connectivity with no guarantees. Although best effort service is the lack of QoS, it provides customers with a reference point on the nondistinct end of the spectrum. Also, best effort is suitable for a wide range of networked applications such as general file transfers and e-mail.

Differentiated Service

In this case, some traffic is treated better than the rest - faster handling, more bandwith on average, lower loss on average. This is a statistical preference, not a hard and fast guarantee. With propoer engineering, differentiated service can be provided and can be appropriate for a wide range of applications. Typically, it is

associated with grouping traffic into a small number of classses, each class receiving a particular QoS in the network.

Guaranteed Service

It means an absolute reservation of network resources for specific traffic. It ussually concentrates on bandwith. It implies reservation of buffer space along with the appropppriate queuing disciplines to ensure that specific traffic gets a specific service level. Bandwith is typically used for the other QoS attributes (jitter, delay) as the widest audiance easily understands it. Bandwith is often reserved down to the level of iindividual traffic flows, and this way partical flows have reserved resources. In other cases, aggregated flows may receive guaranted service.

7.6. Traffic Management

József Wiener, author

Kornél Terplán, reviewer

The term traffic describes the flow of messages through a communications network, whether voice, data or video, analogue or digital. Traffic characteristics are influenced by many traffic metrics.

Traffic management in communications network deals with the controlled use of network resources to prevent the network from having a bottleneck. In particular, when more traffic are allocated to the network resources that they can effectively support, network performance for users degrades. Traffic management controls the user traffic generated by calls entering in and flowing through the network, and prevents the network from overload.

7.6.1. Traffic Control and Congestion Management

The operation by which user traffic is controlled is called flow control. Traffic control should assure that traffic does not saturate the network or exceed the network's capacity. Basically, there are three alternatives for flow control, both for traditional and new (emerging) technologies:

- Explicit flow control: This method limits the user traffic entering the network. The network limits this traffic by introducing an explicit control message, and either the user must stop sending traffic, or his traffic is forbidden to enter.
- Implicit flow control: This technique recommends that the user reduce or stop traffic sending to the network if network traffic exceeds a given value or network situation (e.g. problems in the net) needs contol. One situation can be e.g. that the implicit flow control message is a warning to the user that the user is violating its service level agreement, and overloads the network or part of it.
- No flow control (congestion control): Flow control can be established by just discarding any traffic that is creating problems. This kind of control certainly provides perfect congestion management from the standpoint of the network, but may not meet the performance exceptations of the user.

Based on these principles, different techniques can be used to control (to manage) the traffic. The actually applied techniques depend on the type of the

network and on several other parameters, some of them discussed later in this chapter.

7.6.2. Impact of Different Traffic Shapes

The way of performing the traffic control highly depends on the type of the technology, the network and the service provided on the network. Not so long ago, a phone call was, well, a phone call: people talked to each other. They still do, of course, racking up record POTS traffic, but this goes along with a growing amount of faxes, e-mails, data files, audio-messages, graphics and streaming video. And these newer forms of communications are being delivered over an equally wide selection of protocols — frame relay, ATM, IP, CDPD, GSM, as well as traditional TDM. This growing diversity of media and protocol types—which is (somewhat paradoxically) known as convergence is the governing principle of the broadband, multimedia era.

This diversity means different holding times, speed, sensitivity to delay, etc. i.e. different traffic patterns. For illustration, Figure 7.6.1. illustrates the nature and speed for different type of services and media.

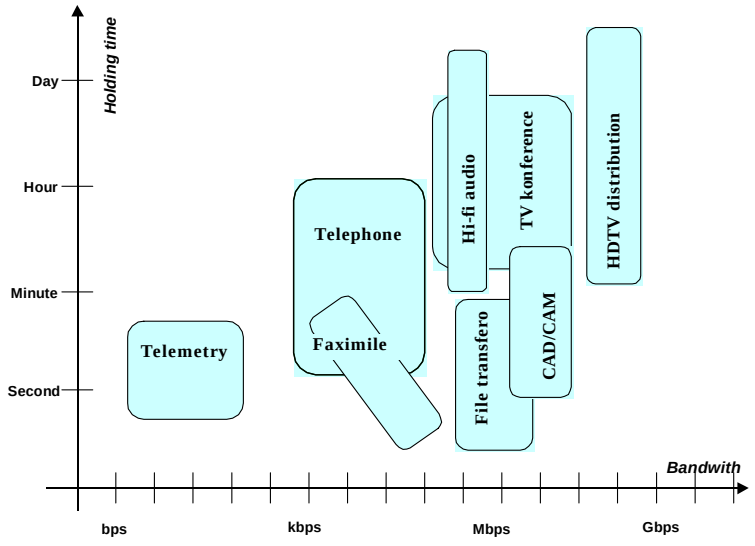


Figure 7.6.1. Traffic Characterisctics for Different Telecomm Services

In the case of voice switches, control is applied at the input of the switch, but in the case of a router it can even be applied in the costumer's router. It can easily be understood that different type of control can be effective at different speeds and

holding times, in case of circuit switched networks and packet switched networks, etc.

7.6.3. Traffic Management in Voice Network

In circuit-switched networks each connection is allocated a fixed amount of bandwidth, and a constant bit rate in the network is provided to communicating entitities throughout the duration of the connection. For example, in a telephone network, each connection requires a 64 kbps channel. If a channel between the caller and the called parties exists, the connection is established, otherwise (e.g. if there is a congestion in the network) it is rejected.

Networks can be overloaded due to an unusually big number of call attemps, or network problems reducing the capacity of the network in some parts. An overload can be

- General network overload, when the entire network is saturated with calls, or
- Focused overload, when only a part of the network is overloaded - e.g. due to a radio-show, or any special event (catastrophe, visit of the pope, etc).

In order to control the network, abnornal conditions shall be detected, and some control methods shall be applied to the network. This action basically can be

- Protective control, when the amount of traffic entering the network is limited (the network is protected against extra load);
- Expansive control, when the extra traffic is re-routed from the source to the destination via areas that is not overloaded, and this way expanding the source-destination capacity of the network.

The theory of this kind of traffic management is well established, there are international (ITU-T) standards available, and there are several practical methods and systems performing such kind of Traffic Management.

The benefits of traffic control in the telephony network is shown by some real examples.

Fig 7.6.2. shows a situation when due to some exchange failures, the completed calls between two exchanges are dramatically reduced. The upper, „floating” curve shows the number of offered traffic (seizures), the „U-like” shows the completed calls. The area between the curves is clearly a loss for the Operator/Service Provider - the smaller the area is, the less the loss is. Without

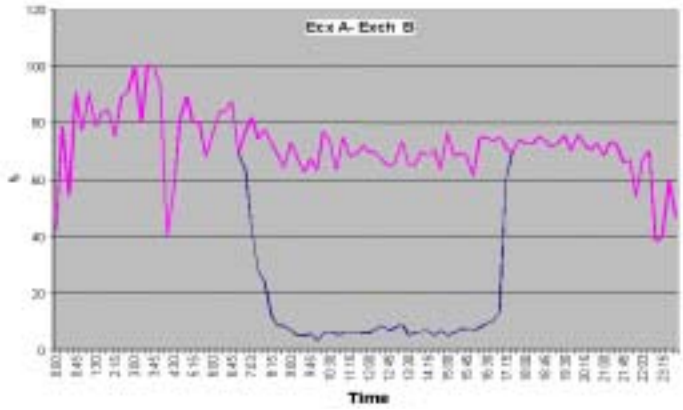


Figure 7.6.2. Traffica Management in Case of Network

Traffic Management, the problem would have been detected later, and the loss in revenue would have been significantly greater.

Figure 7.6.3 shows the impact of solar eclipse in 1999. A huge mass of people was moving to the area where the solar eclipse was best visible. As excepted, people moving to, and beeing there generated an unusually big mobile traffic, that would oveload both the mobile network and the trunks between the wireline and mobile net. In order to avoid the overload the networks, and to maximize the completed calls (maximize revenue), the following actions were introduced:

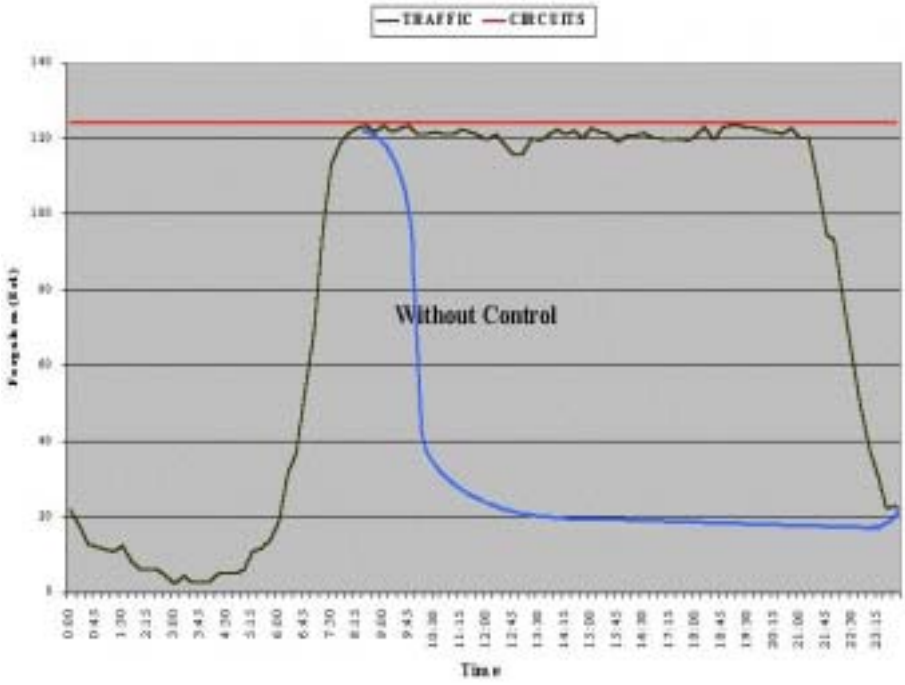


Figure 7.6.3. Traffic Management during the Solar Eclipse in 1999 (Traffic between two controlled exchanges)

- Some routes to and from the area of solar eclipse were rerouted (Expansive Control);
- In some exchanges call restriction control actions (namely in this case leaky bucket control) were applied.

The upper horizontal line on the Figure shows the maximum capacity between two affected exchanges (124 circuits, max 124 Erlangs). The slightly floating curve shows the real carried traffic; it is clearly shown that the applied control kept the traffic almost at a maximum. The dropping curve shows what would have happened if no traffic control had been applied: due to the tremendous amount of unsuccessfull and repeated calls, the real (completed) traffic would have been only part of the maximum capacity. This effect is well known from the traffic theory and also from the practice.

7.6.4. Traffic Management in ATM Networks

ATM networks are expected to support a diverse set of applications with a wide range of characteristics. For the time being, there are no comprehensive measurements to satisfactoraly address the characteristics of all type of traffic. Characteristics of voice sources are well known due to the studies performed in the last century. Constant Bit Rate (CBR) services are relatively easy to manage. The difficulty arises in choosing the (constant) bit rate to provide the desired service quality while minimazing the amount of bit rate used in the network.

Unfortunatelly, the behavior of data sources is not well-understood, and very often unpredictable. There is no typical data application, and no typical source behavior. Image and Variable Bit Rate (VBR) video data sources are relatively new areas, and the knowledge on their source behavior is limited.

In ATM networks, both flow control and congestion control can be used. But, flow control between the source and the destination does not help in reducing the possibility of congestion within the network. To minimize the effects of congestion, each node in the network shall regulate the traffic flow on its input links.

Some aspects of ATM networks that complicate the traffic control problem include the following:

- Various VBR sources generates traffic at significantly different rates - from a few kbps up to Mbps,

- A single source may generate multiple types of traffic with different characteristics,
- ATM networks have to deal with "traditional" performance metrics as call blocking and packet loss, as well as with cell delay variation, maximum delay and skewness,
- Different services have different QoS requirements at considerable varying levels,
- Traffic characteristics of various types of services are not well understood,
- As the transmission speeds increase, the ratio of call duration to cell transmission time increases. There are always a very large number of cells traveling in the network, and the large propagation delays compared with the transmission times, lead to long periods between the onset of the congestion and its detection by the network control elements.
- High transmission speeds limit the available time for on-fly processing.

Considering the high transmission speeds, the control algorithms should be as simple as possible to allow hardware implementations.

Congestion control mechanizms are classified as preventive and reactive control. Preventive control attemps to prevent congestions to occur in the network, The reactive scheme monitors the network for congestion, and if it is detected, sources are requested to slow down until the congestion is cleared.

In practice the efficiency of the control depends on the actual conditions, and a combination of the two principles should be used.

Resource provisioning

This is an important traffic management function for existing networks. Its major role is to provide an acceptable level of connection blocking performance. As the time is going on, several conditions may change in the network. Adding new resources or re-arranging VPs and the bandwidth allocated to them is some kind of traffic management.

Call admission control

When a new connection request is received at the network, the call admission procedure is executed to decide whether to accept or reject the call. Two questions should be answered:

- How can the required bandwith by the new connection be determined?

- How can we make sure that the Service Level of the existing connections are not affected when multiplexing with the new connection?

The technique designed in response to these two questions, should work in real-time and should attempt to maximize the utilization of network resources.

Traffic shaping

For most VBR sources, during the active period, cells are generated at peak rate, while no cells are transmitted during the silent period. By buffering the cells during the peak period before they enter the network, it is possible to reduce the peak rate loading the network. So, the departure rate from the queue is less then the peak arrival rate of cells. Shaping can be done at the source equipment or at the network access point. With this technique good bandwith saving can be achived, but the amount of delay sets limitations.

Traffic policing

Traffic policing or usage parameter control is monitoring the network access points and should detect if the users stay within the connection parameters negotiated during the call setup phase. A policing function should detect a non-confirming source as quickly as possible, and take appropriate actions. When a non-conforming source is detected, the violating cells can be dropped, delayed, marked differently, etc. Most often used control schemes are the Leaky Bucket technique or different Windowing techniques.

Selective discarding

Policing schemes are designed to protect the network from non-conforming users. Hovewer, cells of non-conforming users can be marked, and admitted to the network. This is based on the assumption that these cells can be dropped when congestion occurs. Users may prioritize their cells before they are transmitted to the network. One cell can be either a high-priority or a low-priority cell. Then the buffer space at intermediate points can be used by incoming cells according to their priority.

Reactive congestion control mechanisms

Although preventive techniques reduce the buffer overflow probabilities, it is not possible to eliminate momentary periods of cell losses in the network. Loss cells

result in retransmission by the source codecs, thereby increasing the traffic, and resulting in a catastrophy as turning the momentary buffer overflow periods to sustained periods. Therefore, reactive control mechanisms are also necessary. These mechanisms have been used in low-speed packet-switched networks, but they are not so effective in ATM networks.

The following reactive machanisms can be considered:

- End-to-End Notification Techniques: Once congestion is detected in an intermediate network node, the end nodes need to be notified to be able to react. Then, data sources can reduce their speed to reduce network load. There are three techniques proposed for congestion notification:

  - Estimation by the End Nodes: A source sends time-stamped probe cells along the connection to measure the one-way response time. When the destination node detects congestion (increased response time), it sends a notification to the source node.

  - Explicit Backward Congestion Notification (EBCN): In this scheme, each node in the network monitors the queues of its trunks. When the queue size reaches a predefined threshold, a notification is sent to all sources having path through the congested node.

  - Explicit Forward Congestion Notification (EFCN): Each line monitors the queues, and if its size reaches a threshold, all cells passing through that node is marked with a bit. The receivers know that there is congested node along that connections, but do not react very quickly. Only if the congestion is sustained, is a notification is sent to the sources.

- Adaptive Rate Control: The rate at which traffic is submitted to the network is varied by the source depending on the congestion control information. The proposed scheme is the EFCN to detect congestion.
- Incall Parameter Negotiation: This method is used to minimize the call setup overhead for traffic requering transmission of only one or a few bursts. One alternative is to establish VPs between users of such services and dfining a virtual network for them. In this case, it is relativily easy to manage the traffic inside the VPN.
- Dynamic Source Coding: The source may either reduce the traffic submission rate or mark cells that are not so essential as may be reconstructed from other cells. In case of Dynamic Source Coding, the source may decide which type of control is appliyed in case of congestion, depending on the traffic (delay-sensitive or not delay-sensitive).

7.6.5. Principles of IP Traffic Management

IP traffic management can be considered at various layers of the Internet Protocol stack. Lower layer solutions concentrate on the transmission of traffic from source to destination, but not considering applications and their properties.

At higher levels, the applications or other words their contents should be considered and more intelligent decisions should be taken. Some of the factors affecting the management decisions are:

- The amount of traffic from server to client is much higher than from client to server (asymetric). This suggests that optimization is more important in one direction;
- The medium transfer size of web-documents is small, and short-lived. Due to the bursty nature they are more likely creating congestion;
- There is a difference in the usage of servers. Some top web-servers are accessed much more times than others (focused traffic).
- Some small number of large files consumes a big part of network and server bandwith, and a significant number of small files are accessed only a few times.

Real-time traffic (distance learning, voice, streaming media) is becoming an increasingly proportion of web traffic. Management of this traffic should be supported. An effective management strategy should include admission control, bandwith and buffer allocation solutions.

7.6.6. The Future of Call Control

Call Control Today

Today's call model is basic. The days of the very first call model—picking up a receiver and talking directly to a live operator working a plug-board switch—are long gone. Today, call control is the province of the Intelligent Network (IN). Service-control has been taken out of the switch and put on ancillary SS7 packet-networks to speed the development and deployment of new calling features.

But while these features may be new, they are also narrowband and voice-based. IN was designed expressly for the circuit-switched PSTN of 10 years ago when the Internet was still an R&D project and videos were what you rented from retail outlets. IN and SS7 are built on a triad of Service Switching Points (SSPs),

Service Control Points (SCP) and Signal Transfer Points (STPs), which oversees call completion in countless networks worldwide, aided by Intelligent Peripherals, under the overall control of Service Management Systems (SMSs). SSPs house stored programmed controls that instantly alert a CO's switch that it should consult the SS7 network before proceeding with call routing. Via a series of fast packet STP switches, the SSP then sends a query to the appropriate SCP about how to handle the call. A high-capacity database server, the SCP searches its files for instructions left by the SMS and replies. The switch then executes the forwarded call routine, often calling on an auxiliary IP to play announcements or collect Multifrequency Dial-Tone input from the caller.

In this model, call control is strategically placed in the SSP outright in a software-module called the call control function (CCF) that continually monitors a subscriber's line for points in calls (or PICs), trigger detection points (TDPs) and triggers. These are trip-wires that alert the SSP to temporarily suspend completion of the call.

Changing Call Control

Everything in the telecom world is converging—companies, markets, technologies—everything. It is believed, that for the full potential of this new world of communications to be realized, something else will have to change radically—the conceptual model dictating how calls are set-up and turn-down.

Even though it's the cornerstone of the entire PSTN, the call model in and of itself is changing. For example, in the new IN release, additional triggers - the SSP-IP interface busy and the no-answer trigger - as well as a new category of detection points - EDPs (event detection points)- are added (EDPs are a list of events such as busy, no answer, or terminating available resources.)

A Next Generation Call Model

It is likely that radical new changes will be needed to the call control model if the promise of broadband networking is ever to be fulfilled.

What may be needed is a model that accommodates a series of self-contained mini-calls within a call. The complexities entailed are an order of magnitude greater compared to the present model, requiring a two dimensional call model. Within a year

or two, the market will be demanding for network intelligence of this kind as competitive advantage among broadband service providers increasingly focuses on the ability to offer new services and not on bandwidth exclusively.

Innovative hardware and software vendors focussing on this issue will be rewarded handsomely, but that while each vendor will have a somewhat different solution, they will also all look for guidance for some recent work done by the International Telecommunications Union. What the ITU has done here is to put forth a new, four-plane conceptual map outlining a broadband compliant IN process. In the ITU model, each service is broken down into its constituent parts - generic blocks called SIBs (for Service Independent Blocks) and mapped to the plane immediately below. It is here in the Global Functional Plane, that the Basic Call Process (BCP) resides as a SIB. According to to the traditional IN call model, the BCP defines how calls not requiring any special treatment are handled. Not surprisingly, perhaps, it mirrors its predecessors almost PIC-by-PIC.

The Future of Call Control

The transformation of the old narrowband intelligent network into the broadband intelligent network of the next years will give a somewhat diminished role for the basic call model, which will now be just one building block among many. Instead, based on the ITU model, it will become the service-logic imparted by a string of SIBs and implemented in the distributed functional plane that controls the overall flow of a call through the physical plane. It will be a service logic, in effect, that gives equal footing to both caller-initiated and independent SCP- intitiated calling sub-routines.

However, while intelligent networks are about to undergo radical change, it also indicates that at least the *forms* of the old infrastructure will be preserved. The sub-routines discussed above will continue to be implemented on SSPs (Service Switching Points), STPs (Signal Transfer Points) and SCPs (Service Control Points), but they will be a new breed of SSPs, STPs and SCPs, that vendors will have equipped with considerably more functionality. The venerable SSP will contain a Call Control Function (CCF) that both oversees call processing and provides network connection services; a Service Switching Function (SSF) that activates IN triggers during this call processing; a Specialized Resource Function (SRF) that ensures the

call processing software on a switch can communicate with the service control function on a SCP; and a Call Control Agent Function (CCAF) that supports caller interaction and user access to the network.

Such changes are not likely to be the last iteration of the call control function and that the ITU is not likely to be the only inspiration in this regard. Work is progressing on a uniform IN architecture for fixed-mobile at the European Telephone Standards Institute (ETSI). ETSI has already successfully ported traditional IN Functionality to GSM-based mobile networks via its Customized Applications for the Mobile Network Enhanced Logic (CAMEL) initiative. The International Telecommunications Info rmation Networking Architecture (TINA) Consortium, is designing an architecture that is not backward compatible with the ITU's Intelligent Network Conceptual Model to promote open communications in distributed computing and processing environments, setting the stage for a standards showdown sometime in the near future. This is somewhat troubling as standards disputes have a nasty habit of derailing technological progress.

Probably unlikely that standards disputes could ever really destroy the market opportunities that innovative call control will lead to in the next few years. The huge revenue earning opportunities offered by the next generation intelligent network, means that both vendors and service providers have a very strong economic incentive not to let disputes get in the way of deployment.

7.7. Power supply

Zoltán Janklovics, author

György Varjú, reviewer

7.7.1. Power supply of telecommunication installations and their equipment

Telecommunication installations and equipment planted in them may vary by their design and destination, but they are common in one: all of them inevitably need electric power. In most cases, the public electricity network supplies this energy. In some particular cases it happens that the electric power supply network is not available (for instance in high mountains, far from inhabited areas) for the telecommunication installation. Here, according to the local conditions, alternative energy sources (e.g. solar cells, wind-electric generators, etc.) provide for the feeding of such installations. Due to the fact that in such cases the energy is discontinuously available, these systems are supplemented with interim ancillary energy storing batteries. In Hungary - with the exception of some special cases - the electric energy is available all over the country.

However, the continuity of the electric power supply cannot be always ensured, for a shorter or longer time interruptions may occur due to the outage of the mains supply, or due to transient phenomena, disturbing the operation of the powered equipment. Telecommunications equipment shall work also under such circumstances. During the powering of the telecommunications equipment with energy, the main goal is to provide for the possible most safety and reliable feeding; therefore care shall be taken also for the generation and storage of the energy. The power supply system, consisting of the following system technical elements, performs these tasks [7.7.1]:

- distributing unit of 0,4 kV AC (main distributor),
- stationary emergency (stand-by) generator,
- uninterruptible DC power supply system,

- uninterruptible AC power supply system.

In the past the power supply systems had been usually planted in separate rooms within the telecommunication installations (so called centralised power supply). Nowadays, the accommodation in the same room of the power supply units and the powered equipment is getting more and more widespread (decentralised, or switch-room power supply, however, the stand-by generator is installed even in such cases in a separate room).

System technical configuration of power supply and its system technical units

The central unit of the electric power supply system of telecommunication installations is the 0,4 kV receiving and distributing system (main distributor), to the input interface of which the primary and (eventually) the reserve network feedings, the stationary or mobile stand-by generators are connected. The different consuming appliances of the telecommunication installation are connected to the output of the main distributor. Among them the technological power supply units (uninterrupted DC or AC systems) feeding the telecommunications equipment are of great importance. The other consumers connected to the main distributor can be categorised into two groups: the group of authorised consumers to use Diesel (important consumers, for e.g. climatic appliances, which in case of mains outage, shall be powered from stand-by, emergency generators) and the group of unauthorised consumers to use Diesel.

A) AC distributing unit

The 0,4 kV distributing unit (switchgear), shortly the main distributor, serves for the reception and distribution among the individual consumers of the incoming low voltage (230/400 V) electric power. One characteristic feature of the main distributor of telecommunication installations is that in the interest of increasing the reliability of the feeding, reception of electric power is possible from several directions, i.e. ancillary feeding can also be implemented. In case of mains outage the required AC energy can be assured with the help of stand-by generators. The consumers of the installation can be powered in this case via stationary or mobile stand-by generators. These ones are connected to the main distributor. All consumers of the telecommunication installation get the energy through the main distributor. Therefore each branch has to have over-current protection, switch off possibility and instrumentation, as needed. The main distributor makes the switch over among the

different feeding inputs possible. This can be done manually (by the intervention of the operating staff), or automatically. Against malfunctioning appropriate protections (locking) is built-in into the main distributor. Usually the quantity of the energy consumed from the public electric network is metered also in the main distributor.

Basically there are two system technical versions of the main distributor known: a design with one bus and with double buses. The more complicated two-bus system is so arranged that each feeding and each branch can be connected to both of the buses, so thus the required variation of connection can be created. This arrangement makes the maintenance and repair of the main distributor, the replacement of the components easier and improves the operational safety. The price of this flexibility of such a two-bus system is that from each switching element, bus, or cable double quantity is needed. Accordingly, the space required by the main distributor and the demanded maintenance, as well as the investment costs of the system significantly increase. Both the one- and two-bus systems are widely used in the practice. These main system technical designs have several versions. For instance, the flexibility of one-bus systems can also be increased significantly using longitudinal bus sectioning.

The size of the main distributor, the number of branches the loadability of the used switching and protecting components always depend on the requirements of the telecommunication installation. In general the main distributor is a unit of individual design and installation. In small installations, which do not need operating staff, the main distributor is a small wall cabinet. In larger installations the powerful distributors, constituting an individual group of equipment, are installed either in separate room or commonly with the stand-by generator (or eventually with the technological power supply). The advanced types of main distributors are of modular-rack design. The individual cabinets have racks for the bus, the cable and the units. An important requirement against the main distributor is that the switching configuration shall always be transparent and the most important parameters can be read on the spot. In case of automatic-operated main distributors, reasonably the possibility of manual switching operations or manual control shall be executable in case of failure of the automatic system. Special attention shall be paid to reliability at the design of the main distributors because we have in vain alternative, stand-by

feeding input, emergency generator, if - due to an unreliably functioning switching element - these units cannot feed the equipment.

B.) Uninterruptible DC power supply system

In a telecommunication installation the main consumer, requiring usually the most energy supplied, is the telephone exchange. For the supplying DC voltage is required, having usually a nominal voltage of -48 V_{dc}. The most important requirement for the supply is the continuity, i.e. the powering system must not interrupt even for a second the supplying of the DC voltage. Otherwise the built-up connections would disconnect (in case of the exchanges it would necessitate the re-programming of the exchange, which in the given case could block the provisioning of telecommunication services for hours in a region). Rectifiers fed from the public mains produce the required DC voltage. However, these units can only provide the DC voltage, if the required mains voltage is available. Since the continuity of DC current supplying is to be ensured, and the uninterruptible supply cannot be assured only from the AC current side, the DC current needs to be directly stored. The most suitable for this purpose are the rechargeable and dischargeable chemical energy sources, the batteries. The uninterrupted DC systems are so designed that in case of failure of the mains voltage, the batteries can immediately replace the falling out DC voltage of the rectifiers. The simplest method to implement this is the following: the proper battery set is connected parallel to the rectifier providing the DC voltage, so thus the rectifier - beside the feeding of the consumer - is charging the battery as well (parallel floating charging mode of operation). In case of a mains outage the battery set is taking over immediately (without interruption) the feeding of the consumer. The lead acid batteries can be used for this type of operation. Considering that both in case of DC or AC power supply systems, the key component of uninterruptible operation is the battery, in the followings the batteries applied in the mentioned systems will be discussed in details.

Lead acid batteries can be categorised into the following main *two groups: flooded electrolyte batteries* and gel technology *batteries*. In the first type diluted sulphuric acid in liquid state is used as electrolyte. Because of the evaporation of the sulphuric acid these types of batteries need to be accommodated in battery rooms designed specially for this purpose, and beside the charging they require maintenance (e.g. liquid level control), too. In case of gel technology batteries the

sulphuric acid is either in jelly state or in a certain intermediate insulator. This makes it possible to produce these batteries in closed form, i.e. in this case the electrolyte has no contact with the air space. There is a safety valve mounted onto the housing of the battery, which has the task to release the overpressure developed within the accumulator in case of disturbance in operation. The gel technology batteries are more sensitive to the charging voltage and to the ambient temperature. (According to the abbreviation coming from their English denomination, these batteries are called also as VRLA batteries.) The VRLA (Valve Regulated Lead Acid) batteries may be accommodated also in common room with the telecommunications equipment, for e.g. in the frame of it.

Regarding the lead acid batteries the following (cell) voltage levels can be mentioned:

- Nominal voltage: 2V.
- Floating charging voltage: The charged level of the battery can be maintained with floating charging. The value of it is 2,23V±1%.
- Boost charging voltage: The discharged battery can be recharged with the help of the boost charging. The value of it, which is limited for e.g. by the developing gas, is usually 2,35-2,4V, but sometimes the manufacturers allow higher values.
- Discharge final voltage: below this voltage the processes going on in the battery become irreversible and the battery deteriorates. The value of this voltage is ca. 1,8V.

We mention that VRLA battery sets require temperature-dependent charging voltage, and the boost charging is usually not allowed in their case. The time, during which the telecommunication system can be powered without the public mains voltage, depends on the capacity of the battery. This time is called as back-up time. In the interest of continuous operation of the telecommunication system the public mains voltage is to be replaced within the back-up time. In case if the voltage tolerance of the consumer requires it a more complicated system shall be applied instead of the parallel floating charging system (e.g. consisting of a 24-cell battery set and of a rectifier operating parallel with it). (Such complex systems can be, for instance, a dropping diode system, or serial converter system, which are used in analogue exchanges with narrow voltage tolerance and in certain types of digital exchanges.)

For the proper operation of the given telecommunications equipment, appropriate power supply is to be ensured. The parameters that are to be met by the power supply, can be specified for every equipment. The most important such parameters are:

- limit values of the output DC voltage of the power supply system;
- power consumption of the supplied equipment;
- the still tolerable maximum value of the noise voltage superimposed onto the output DC voltage (the noise voltage requirements mean one of the quality distinctions between the power supply systems of telecommunications equipment and the power supply systems of other purpose.);
- the maximum value of output voltage changes (transients) caused by the change of load or mode of operation.

The above mentioned are the requirements set by the powered equipment against the power supply. The energetic aspects of power supply (e.g. efficiency, inrush current) are worth to be mentioned and the EMC requirements of the relevant standard specifications shall be met, as well [7.7.4].

C). Uninterruptible AC power supply systems

In telecommunication installations including telephone exchange, the technological equipment have required up to the near past mostly direct current power supply. Nowadays, those, first of all IT type telecommunications equipment, are spreading quickly, which require power supply with 230V alternating current. These systems are also sensitive to the continuity of the power supply. The uninterruptible alternating current systems serve for the satisfaction of these demands.

The demand for the uninterruptible power supply (UPS) requires the intermediate storage with batteries of the energy in this case, as well. At the same time, a unit (inverter), which transforms the stored DC voltage into one- or three-phase alternating voltage, having the frequency and (sinusoid) waveform of the mains current, is also needed. The inverter is the determinant key element of the uninterruptible alternating system. In the case, when the demand for the uninterrupted alternating current is less than the DC output power demand, the (48V) battery of the DC system can be used for the energy storage.

Sensitive consumers are fed from continuously operating inverters (inverter-based operation mode). If mains voltage is available, the inverter is working

synchronised to the mains. (So thus, there is no voltage jump occurring in the feeding voltage in case of a mains outage.) In battery-operated mode the inverter is self-running. In case of failure of the inverter, or if it is switched off, the role of the inverter is taken over by the mains. Even in such cases the switchover must not result-in any voltage jump, therefore a high-speed electronical switch (usually a Thyristored solution) controlled by an electronical device is applied (electrical by-pass). The systems have also the possibility of mechanical switchover (mechanical by-pass), which might be necessary for service purposes. This switching device is equipped with appropriate locking. (The output voltage of the inverter must not be interconnected with the mains.)

If the required power is comparable with the output power of the DC system, the application of 48V batteries would not be economical (the capacity of the cells should be increased considerably). Instead of it, the system itself is equipped with an own battery set and charging unit. The battery sets of the alternating current systems are of higher voltage (of 200...400V), which means a higher number of cells (100...200). The higher voltage makes it possible that the same power can be stored with the help of lower capacity cells, which is a more economical solution than using of battery sets having less cells but higher Ah capacity.

D). Stand-by (emergency) generators

Key important consumers are supplied with electrical energy from stand-by (emergency) power generators in case of outage of the public mains over a certain period of time. The stand-by generator itself is basically a small-sized electrical power plant, which consists of an alternating generator driven by an internal-combustion engine. The stand-by generators used widely in the practice do produce 0,4 kV three-phase alternating current and a diesel motor is used in them as power engine. In telecommunication installations there are two different types as to the design of such generators are applied: larger, important installations have built-in – stationary (fixed) – stand-by generators, while to smaller systems movable/transportable – i.e. mobile – stand-by generators can be connected.

The engine part of the stand-by generators of today is a turbo-diesel motor, made of light alloy. Alternatively, gas fuelled engine can also be applied. For the quick start and loadability of the motor, the cylinders of it are preliminarily heated (in order to keep the temperature) with the help of the cooling water. The automatics

provides for the start of the motor, the regulation of the speed (r.p.m.) and the fuel injection, etc. In addition, it has other important function, too: to indicate and forward to the monitoring system the functional problems, signs of malfunctioning. In case of an outage of the mains, the automatics start the stand-by generator (usually with a 1-2 minutes delay-time). The generator is designed with alternating, three-phase voltage.

Stationary, fixed stand-by generators are often co-located with the main distribution unit. This goes along with that the drainage of the exhausted gases, as well as the sound-proofing of the room must be solved. A separate battery set, having its own charger, starts the diesel motor. The construction of the main distributor and that of the stand-by generator must be in line with each other, since the benefits of the automatic-start stand-by generator can only be exploited with an automatic main distributor.

The construction and design of mobile stand-by generators is basically identical with that of the stationary types.

7.7.2. Powering of wire communication networks

In the traditional wire line telecommunication systems the individual customers are connected via metallic conductors (cables) to the nearest telephone exchange. The metallic conductor (e.g. copper pair) is used for telecommunications and for the transmission of signals necessary for the building up of the connections. The required energy is provided by the powering system installed within the exchanges, or the energy is supplied via the metallic network to the equipment to be powered.

In this section a short overview is given for the powering solutions of wire line networks different from the traditional ones.

Similar powering solutions are applied in the optical access networks and in metallic wire digital networks. (These solutions will be described as an example of optical access networks.)

A.) Power supply of optical access networks

In case of optical cables there is no metallic contact between the two endpoints of the cable. The connected equipment includes electrical circuits, which

require electrical energy for their operation. In case of equipment installed in telecommunication systems, the powering is solved: the uninterruptible (48V, direct voltage) power supply systems of the telephone exchanges – beside the powering of other transmission equipment – are feeding the equipment of the optical systems, too.

The uninterruptible powering of equipment installed in outdoor (street) cabinets, in the customer premises or located in the vicinity to the customers, requires individual solutions. The active elements of the optical network can be found in geographically different places, which sometimes are quite far from each other.

In case of remote powering the equipment receive the energy necessary for their operation either from the nearest uninterruptible power supply, or from the power supply system (Power Node, PN) of one other active element of the network. The precondition of remote powering is that the remote power unit (RPU) and the powered equipment shall be interconnected also with a metallic (remote powering) cable.

Accordingly, there are two major versions of remote powering:

- centralised powering, whereas the equipment are supplied with power through the Remote Power Unit, located within the telecommunication installation, and
- cluster powering, whereas Power Nodes (PN) belonging to the equipment (e.g. street cabinets) installed within the network are powering the subscribers, or the network units (ONUs), which do not have own powering system, but require powering.

The remote power voltage is supplied by the Remote Power Units (RPU), the output circuits of which are secondary circuits.

The values of the voltages applied in telecommunication systems can be categorised from safety aspects according to standards [7.7.5 -7.7.6] into the following groups:

- Safety Extra Low Voltage (SELV) circuit: a circuit, where under normal operating conditions the direct current shall not exceed the value of $60V_{dc}$, or the alternating current the 42,4 V_{ac} value. Moreover, in the event of a single failure the voltage shall not exceed the highest values given for normal operating voltage, for longer than 0,2 s. Moreover, the limit of $120V_{dc}$ or 71 V_{ac} shall not be exceeded.
- TNV1 Telecommunication Network Voltage circuit: a circuit, which complies with the requirements specified for safety extra low voltage circuits, but which is subjected to overvoltage from the telecommunication network.

- TNV2 Telecommunication Network Voltage circuit: a circuit, the normal operating voltage of which exceeds the limit value specified for safety extra low voltage circuits, and which is not subjected to overvoltage from the telecommunication network.
- TNV3 Telecommunication Network Voltage circuit: a circuit, the normal operating voltage of which exceeds the limit value specified for safety extra low voltage circuits, and which is subjected to overvoltage from the telecommunication network.
- Limited current circuit (a circuit limiting the remote feeding current): a circuit, which even in the case of breakdown of the insulation, or defect of any element, complies with the following requirements:

  - For frequencies not exceeding 1kHz, the steady-state current drawn through a resistor of 2000 Ω connected between any two points, shall not exceed the value of 0,7 mA AC or 2 mA DC.

  - For (accessible) parts not exceeding 450V peak AC or DC, the circuit capacitance shall not exceed 0,1 μF.

  - For (accessible) parts exceeding 450V peak AC or DC, but not exceeding 15000V peak or DC, the (available) stored charge shall not exceed 45 μC.

  - For (accessible) parts exceeding 450V peak AC or DC, but not exceeding 15000V peak or DC, the available energy shall not exceed 350 mJ.

In case of *centralised powering*, the primary side of the RPN is connected to the 48V power supply system. So thus, the continuity of the remote powering is also assured. Such solution is applied, for instance, at ISDN circuits, or at PCM systems.

In case of *cluster powering* the network unit providing the remote powering shall have its own power supply system. The input energy source of it is the public mains. A continuous (uninterrupted) powering is to be assured also in the event of outage of the public mains; therefore the power supply system shall include battery sets as well. The power supply system, besides the remote powering, has to feed with energy the locally installed equipment, as well; therefore it often includes several consumer outputs of different voltage levels. The battery voltage may be also different from 48V.

In case of *local powering*, the primary energy source of the equipment is the public mains. Since the equipment need to operate also in the event of outage of the mains, an uninterruptible power supply system shall be applied. The construction and the design of the power supply depend on the place of installation of the equipment and on that if it serves one or more subscriber(s).

At the installation of the local powering system, the following issues shall be dealt with:

- the connection to the energy supply network,
- the installation environment of the equipment (street cabinet, etc.),
- the selection and accommodation of the battery set,
- the back-up time of the battery set.

The majority of the equipment of the optical network can be found not in air-conditioned buildings, but these are installed in street cabinets, staircases, etc. Therefore, special care shall be taken for the observation of environmental conditions of the selection and installation of power supply equipment. (For instance, the applicable operating temperature range is very important to avoid overheating problems.)

The selection and installation of battery sets require special attention, too. It is important that only maintenance-free batteries can be built in into the equipment of the optical networks. In customer premises equipment, besides the VRLA-type battery sets Ni-Cd batteries can also be applied. At the same time, for VRLA batteries the recommended operating temperature range (usually it is around 20 °C), shall be taken into account, because at a higher temperature the life-time drastically decreases, or at a lower temperature the full charged quantum cannot be taken out of the battery. The lower temperature range is also limited. Therefore, it is necessary to ensure somehow (e.g. by putting the batteries into a man-hole, or cooling/heating the cabinets) that also the batteries of street cabinets operate under equalized temperature conditions.

While power supply systems of telecommunication installations are capable to bridge over almost without limitation any mains outage, in case of ONUs, which are operating with local powering, the bridge-over time is strictly equal to the back-up time of the built-in batteries (except mobile Diesel operation). So, it is important to choose properly the back-up time of the batteries. It depends on the reliability of the local electrical supply network. When making our choice, the frequency and the average duration of mains outages (failures) occurring in the given area, shall be taken into consideration. In larger cities the electrical network is more reliable, because (in most cases) the supply of energy to the consumers is realized via from the weather protected, buried cables and the distances are smaller (the fault repair

times are shorter). The situation is worst in small, remote settlements and scattered farms, i.e. in places where both the distribution and consumer's networks are made of overhead cables and the consumers are far from the maintenance and repair staff. Therefore, it is reasonable to apply batteries of different back-up times for the equipment, depending on whether they are installed in larger/smaller cities, villages or settlements of scattered farms.

B.) Powering of CATV networks

Cable television (CATV) networks are of hierarchical topology that can be segmented to several individual parts and according to this segmentation also the applied method of powering is different. In general, in CATV networks there are coaxial cables used, but the network may consist of optical fibre cables as well. Beside the passive elements, such networks include line amplifiers and house amplifiers, too. Remote powering usually does the energy supply of line amplifiers. The feeding of house amplifiers can be solved both with local and remote powering. The powering of the CATV head-end stations is to be mentioned, as well, whereas it is always to be constructed and designed according to the local requirements. Head-end stations are always locally powered. The individual equipment may require either 230V AC or 48V DC powering. In case of head-end stations installed in telecommunication systems the uninterruptible DC voltage powering of the station can be solved via the $48V_{dc}$ power supply system of the installation. If uninterruptible AC voltage supply is required - and if in the installation an uninterruptible alternating current network is not available - the powering can be solved also with the help of individual uninterruptible power supply units (UPS).

Remote powering of CATV amplifiers: Amplifiers, which are applied in the trunk (or main line), or line and distribution network layers of the CATV network, are remotely powered with alternating voltage (maximum 60V), transmitted usually via the coaxial cables. The house amplifiers used in in-house networks are locally powered.

The applicable supply voltage: Because of the risk of corrosion, direct voltage is practically not used in CATV networks for remote powering. Taking into account the 24 V stabilised internal supply voltage applied for the powering of amplifier modules, as well as with consideration to the electrical safety specifications, max. 60V alternating voltage is applied for remote powering. The built-in powering units of

the amplifiers convert this 60V alternating voltage into stable direct voltage. The lower limit of the input alternating voltage is between 27...40V depending on the construction of the amplifier. The remote supply voltage is (usually) produced with the help of ferroresonant voltage stabilisers. At the output of it, a nearly quadratic (trapezoid)-form output voltage appears, which is excellent for the reduction of the output voltage fluctuations, occurring due to the loading changes, or the input voltage fluctuations.

Ferroresonant power units: The alternating voltage stabiliser operating on the principle of ferroresonance, is a special transformer, in which the stabilising effect is achieved by the driving of the iron core into a saturated state. For doing so, beside the usual primary and secondary coils of the transformer, also a third coil is necessary, to which in parallel a capacitor is connected. The third coil along with the capacitor forms a resonant circuit tuned to the network frequency (50 Hz). The stabilising effect is reached by getting the iron core into a saturated state, before the sinusoid input voltage would reach its peak value, so thus, on the output side the voltage practically cannot change. Similarly, even under the effect of changes of the load, the amplitude of the output voltage shall not remarkably change, either. The 'price of it' is that the output voltage is not sinusoid, but it has a trapezoid waveform. The advantage of the ferroresonant stabilisers derives form their relative simplicity and consequently in their reliability and cheapness. Therefore, such power units are greatly applied in CATV networks. The relatively high inrush current (which among others depends on the pre-magnetic status of the iron core, i.e. on the moment of time of the previous switch-off) is a disadvantage of such units. Therefore, the circuit breakers providing for the over-current protection on the input side, shall be selected with due care. The efficiency of the unit reaches a 90% level approximately at half-load.

7.7.3. Power supply interface of telecommunications equipment

One basic condition of the implementation of a power supply system, which is optimal both from technical and economical points of view, is that the parameters of the power supply interfaces of the powered equipment shall be standardised or uniformed. (The power supply interface is the common interface between two

functional units, at which the power unit is connected to the telecommunications equipment.) With the standardisation of the parameters of the power supply interface the compatibility between the powering and the powered equipment can be ensured. At the same time, the conditions necessary for the normal operation of the equipment can be maintained also in the event, when the power supply equipment is feeding several, different telecommunication systems. The standardisation of the parameters makes it possible that telecommunications equipment to be implemented newly can (on a longer term) be connected to already existing power supply systems, without the application of special adapters (e.g. DC-DC converters). Moreover, it will be also possible that power supply units having similar parameters can be applied in all telecommunications equipment. The parameters and characteristics of power supply interfaces are specified in the series of Standards ETSI 300 132 [7.7.2 - 7.7.3]. (The requirements are related to telecommunications equipment operated by 230V alternating current, or by 48V direct current. There are no international specifications available for systems differing from these ones (e.g. for 27 V direct current systems). Similarly, the remote power supply systems are not standardised either.)

Requirements of power supply interface operated by alternating current (AC)

According to Standard ETS 300 132-1, the following major requirements shall apply to power supply interface(s) of equipment powered by UPS of 230V AC nominal voltage:

- The supply voltage range shall be 207-253V (between the neutral and the other conductor); the frequency range shall be 48-52 Hz.
- The equipment shall not suffer any damage (either in its hardware or in its software), if the value of the supply voltage is between 0 - 207V. The frequency range in this case may be between 45 - 55 Hz.
- Following the restoration of the supply to the normal (operating) voltage range, the equipment shall then resume the operation according to its specifications, without requiring any intervention.
- Without the change of its operating parameters, the telecommunications equipment shall be capable to bear voltage fluctuations, occurring within the specified limits due to the regulation of the secondary voltage.
- Among others, the equipment shall also comply with the requirements specified for the inrush current, the harmonic content of the input current, the surge voltage (over-voltage) immunity and the immunity to radio frequency disturbances.

Requirements of power supply interface operated by direct current (DC)

According to Standard ETS 300 132-2, the most important parameters relating to power supply interface in case of equipment requiring direct current power supply, can be summarised as follows:

- The normal service voltage range for the –48 V direct voltage nominal supply at the interface shall be -40,5 to –57 V.
- Telecommunications equipment operated at –48 V nominal direct voltage shall not suffer any damage when subjected to the following - from the nominal value different - voltage ranges: 0 to -40,5 V DC and -57,0 to –60 V. Following the restoration of the normal voltage range, the system shall continue to function within its operational specification without requiring manual intervention.
- The equipment shall comply with the requirements specified in the above standard for transients, voltage drops, as well as for weighted and wideband noise immunity and emission not exceeding 20 kHz. (The requirements for values above 20 kHz are specified in EMC standards.)

List of abbreviations:

Abbreviation:	English equivalent:	Hungarian equivalent:
ONU	Optical Network Unit	optikai hálózati egység
PN	Power Node	tápellátó csomópont
RPU	Remote Power Unit	távtápláló egység
SELV	Safety Extra Low Voltage	biztonsági feszültség
TNV	Telecommunication Network Voltage	távközlési hálózati feszültség
UPS	Uninterruptible Power Supply	Szünetmentes tápellátó rendszer

References

[7.7.1] Janklovics Z. - Gerdai G.: Telecommunications power supply; Magyar Távközlés, 1997. August (in Hungarian)

[7.7.2] ETS 300 132-1 Equipment Engineering; Power supply interface at the input to telecommunications equipment; Part 1: Operated by alternating current (ac) derived from direct current (dc) sources.

[7.7.3] ETS 300 132-2 Equipment Engineering; Power supply interface at the input to telecommunications equipment; Part 2: Operated by direct current (dc).

[7.7.4] ETSI EN 300 386 Electromagnetic compatibility and Radio spectrum Matters (ERM); Telecommunication network equipment; Electromagnetic Compatibility (EMC) requirements.

[7.7.5] EN 60950 Safety of information technology equipment, including electrical business equipment

[7.7.6] EN 41003 Particular safety requirements for equipment to be connected to telecommunication networks

7.8. Electromagnetic compatibility

Zoltán Janklovics, Ferenc Lénárt, author

György Varjú, reviewer

7.8.1. EMC fundamental definitions, terms and principles

In the interest of proper functioning of equipment in the given environment, i.e. without the degradation of its performance, it is necessary that the different equipment and their operational environment are made electromagnetically compatible. *The electromagnetic compatibility*, or the commonly used abbreviation, EMC (according to the English term), refers to such a professional field, *the object of which is to eliminate or at least to possibly minimise the „mismatch" between the equipment and their operational environment* according to the accepted norms, standards and regulations. In the EMC Chapter of the International Electrotechnical Vocabulary [7.8.1], the term of EMC is defined as follows: „The ability of an equipment or system to function satisfactorily in its electromagnetic environment without introducing intolerable electromagnetic disturbances to anything in that environment".

The electromagnetic interference (EMI) means degradation of the performance of a device, equipment or system caused by an electromagnetic disturbance. Electromagnetic disturbance is any such electromagnetic phenomenon, which degrades the performance of a device, equipment or system, or adversely affects living or inert matter. The EMI phenomenon, the occurring degradation of performance consists of three ingredients, namely:

An emitter, i.e. a source emitting the electromagnetic disturbance;

- A susceptor, i.e. a susceptible (disturbed) device, equipment or system showing degradation of its performance;
- A medium in between, which is called the coupling path.

The definition of the term degradation is as follows: An undesired departure in the operational performance of any device, equipment or system from its intended performance.

It is important that in the definition the adjective „undesired” is used and not the adjective „any”. Therefore, for the tests the kind of undesired departure in the operational performance must be clearly specified (see performance criteria).

In the everyday practice there are usually lots of artificial or natural sources, which are emitting electromagnetic disturbance, creating such electromagnetic environment, in which potential susceptors can be found. Due to the great variety of the possible situations, the electromagnetic environment is very complicated.

EMI has relation to EMC from the two following aspects:

1. The emission determines that condition, under which a given electrical or electronic system is functioning properly, without introducing electromagnetic disturbances, causing degradation of performance in other systems;

2. The immunity determines that condition, under which a given electrical or electronic system is capable to function properly in the given electromagnetic environment, without the risk of degradation of performance.

With some simplification the basic principle of assurance of electromagnetic compatibility can be illustrated as follows. Figure 7.8.1 shows a possible combination of an emission and an immunity level and their associated limits as a function of an independent variable quantity (for e.g. the frequency) in case of single disturbance source (emitter) and a single susceptor.

Limits and levels for a single emitter and susceptor as a function of some independent variable

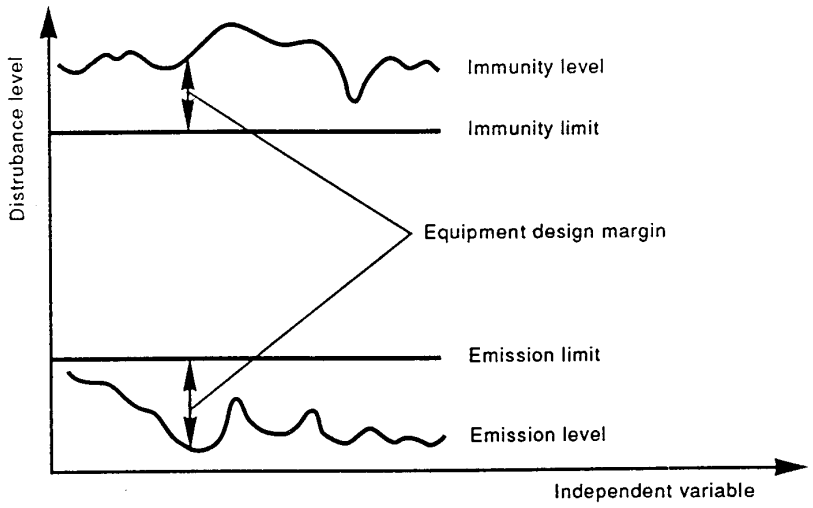


Figure 7.8.1.

According to Figure 7.8.1 the emission level is always lower than its maximum permissible level and the immunity level is always higher than its minimum required level, i.e. the immunity limit. Hence, the emitter and the susceptor comply with the requirements. In the figure there is some margin between the measured level and the limit value. This margin might be called as “equipment design margin”, and is an additional margin in the design to ensure compliance with the limit and it is controlled during the EMC test. The range between the immunity limit and the emission limit is the compatibility range. The compatibility level is within this range.

If the emission level and the immunity level have been designed in a way that they meet the existing electromagnetic phenomenon, then Figure 7.8.1 indicates an electromagnetically compatible situation. The figure shows that the immunity level is higher than the immunity limit and this is higher than the emission limit, which, in turn, is higher than the emission level. [7.8.2].

The two key issues in the problem are the emission and the immunity; the harmonisation of these two criteria is a regulatory task.

7.8.2. Categorisation of EMC according to phenomena

The following table gives an overview of the principal electromagnetic disturbance, which shall be considered accordingly to the list commonly agreed by the IEC and CENELEC [7.8.18].

Note: There is of course no abrupt limit between the low frequency domain and the high frequency domain but a soft transition between 9 kHz and 150 kHz. For formal applications the limit is set at 9 kHz.

The above disturbance phenomena can be classified into the following groups:

According to frequency:

- Low-frequency (less than 9 kHz) disturbances;
- High-frequency (more than 9 kHz) disturbances.

According to mode of propagation (coupling) of the disturbance:

- Conducted disturbances;
- Radiated disturbances.

Conducted low frequency phenomena	Harmonics, interharmonics Signalling voltages Voltage fluctuations Voltage dips and interruptions Voltage unbalance Power frequency variations Induced low frequency voltages d.c. in a.c. networks
Radiated low frequency field phenomena	Magnetic fields[a] Electrical fields
Conducted high frequency phenomena	Directly coupled or induced continuous voltages or currents Unidirectional transients[b] Oscillatory transients[b]
Radiated high frequency field phenomena	Magnetic fields Electrical fields Electromagnetic fields continuous waves transients[c]
Electrostatic discharge phenomena (ESD)	
High altitude electromagnetic pulse (HEMP)[d]	
[a] Continuous or transients **[b] Single or repetitive (burst)** **[c] Single or repetitive** **[d] To be considered under special conditions**	

In addition the disturbances can be differentiated also according to their duration in time as:

- Continuous
- Short time (transient) phenomena.

Special phenomena:

- Electrostatic discharge (ESD, combined – conducted and radiated phenomena);
- High Amplitude Electromagnetic Pulse (HEMP), or Nuclear Electromagnetic Pulse (NEMP)
- Lightning Electromagnetic Pulse (LEMP).

The standards are ranking the impacts of electromagnetic pulses of lightning among the conducted and radiated phenomena, while the phenomena are categorised rather according to their physical characteristics, than on the basis of their sources (ESD and HEMP phenomena are exceptions from this point of view).

7.8.3. EMC regulation

In the developed industrial countries the EMC constitutes part of both the technical and legal, statutory regulation. Within the frame of it, the state(s) stipulate(s) mandatory technical requirements for the products and forms of conduct and behaviour, which ensure the observation of these stipulations. Technical regulation is adopted by the state(s) only in cases when a certain defective product is greatly hazardous from the point of view of protection of life, health, environment and property.

The statutes lay down requirements and compliance for manufacturers and distributors. At the same time, there are stipulations given in these statutes also for the ways and means of confirmation and declaration of compliance.

In order to promote unification in the field of EMC regulation in the European Union, the Directive No. 89/336/EEC on Electromagnetic compatibility was issued on 3 May, 1989 (hereunder referred to as: EMC Directive) [7.8.3]. The EMC Directive, which entered into force in the EU in 1996, specifies the scope of its application, as well as those institutional, personal, standardisation, quality certification, statutory, etc. conditions and requirements, the fulfilment of which is required to the conformity with the EMC requirements and in case of products to obtaining the CE mark of conformity. It is worth to mention that an exact interpretation of the EMC Directive is not simple. To help it, the EU has published a guiding document [7.8.4].

The Hungarian EMC Decree is based on the 89/336/EEC Directive and it was issued as 31/1999 (VI.11.) GM-KHVM Common Decree on electromagnetic compatibility and entered into force in 1999 [7.8.5]. (The up-to-date list of endorsed national standards referred to in the Decree is regularly published as Annex to the Hungarian Standardisation Bulletin (Szabványügyi Közlöny). Remarkable that the first list contained 96 standards, which also reflects the complexity of the subject.)

On one part, the importance of the Hungarian EMC Decree consists in that that it stipulates uniformed protection requirements both in respect of emission and immunity in the whole frequency range within the whole scope of EMC, on the other part, that it is in line with the legal harmonisation principle of the European Union and it complies also with the EMC standards of the Community.

Within the scope of this field the Decree specifies the protection requirements and the controlling tasks. It stipulates that „…the equipment shall be so constructed that:

a) the electromagnetic disturbance it generates does not exceed a level allowing radio and telecommunications equipment and other apparatus to operate as intended;

b) the apparatus has an adequate level of intrinsic immunity to electromagnetic disturbance to enable it to operate as intended."

In addition, as one of the most important protection requirements, it stipulates that the equipment should be constructed in such a way that it has an adequate level of electromagnetic immunity in the usual electromagnetic environment where the apparatus is intended to work so as to allow its unhindered operation, taking into account the levels of disturbance generated by apparatus complying with the standards pertaining to the Decree.

The laying down of the adequate system of EMC standards is not the only, but no doubt, one of the most important preconditions of the application of the EMC Directive.

The *system of EMC standards* consists of the following three types of standards [7.8.6]:

a) Basic standards give terms and definitions regarding the disturbing phenomena and the description of the phenomena, the detailed testing and measurement methods, the testing devices and measurement configurations of basic tests.

The list of harmonised standards does not contain the basic standards, since from the point of view of declaration of the conformity of products not the basic standards are prevailing.

(For instance the IEC 1000-4-X series contain basic standards, which comply with the above criteria.)

b) Generic standards give the EMC requirements (limits), as well as the standardised testing methods that can be applied to equipment operating in a given environment.

c) Product family standards give detailed specifications for emission and immunity requirements of equipment in question.

With regard to the complexity of the system of EMC standards, we give a short overview about the present practice of marking of standards.

The IEC (International Electrotechnical Commission) has introduced the marking of IEC 1000-X-Y for the series of EMC standards. According to the second subdivision (X) of these series, the subjects (- following the marking and numbering structure of IEC -) are:

1. General considerations, definitions, terminology
2. Description and classification of the environment, compatibility levels
3. Emission limits, immunity limits (in so far as they are not regulated in a product family standard)
4. Testing and measurement techniques (Basic publication)
5. Installation and mitigation guidelines
6. Generic standards of series numbered as 1000 and 61000
7. Miscellaneous

European Standards with numbering EN 61000-X-Y are those standards, which have entered into force with the adaptation of the relevant IEC 1000-X-Y numbered standards of IEC.

(Recently, IEC is also using the numbering structure of IEC 61000-X-Y, that means that the standards corresponding to each other, issued by the different committees, vary only according to their letter codes.)

EN 60XXX series number have those CENELEC standards that have been adapted from IEC but not belong to the IEC 1000 series. (For instance, hereto belong the EN 60555 series of standards, which deal with disturbances caused in power networks by household appliances and similar equipment.)

The European standards, dealing with radio frequency disturbances, have been taken over from CISPR documents and they have numbering of EN 55XXX series. (Hereto belong for e.g. the emission-related product family standards numbered as EN 55011, EN 55013… EN 55022 (corresponding to CISPR 11, 13, 14, 15 and 22 requirements).

The EMC standards developed by CENELEC are standards numbered as EN 50XXX series. Such general publications are the emission (EN 50081-1, 2) and immunity (EN 50082-1, 2) standards issued on residential, commercial and light industry environment. (These are continuously transferred to general standards of series 61000-6-Y.)

The majority of the telecommunications related standards have been elaborated by ETSI. These are published either as ETSI standards with numbering ETS 300… or as CENELEC standards with numbering EN 300…. The latter ones are, first of all, harmonised standards under the EMC Directive.

7.8.4. Requirements

The majority of *equipment-related requirements* has been produced by ETSI Technical Committee (Electromagnetic Compatibility and Radio Spectrum Matters – ERM). The requirements applicable for the considerable part of telecommunication network equipment are defined in the harmonised standard EN 300 386-2 (Electromagnetic Compatibility and Radio Spectrum Matters; Telecommunication network equipment, EMC requirements; Part 2: Product Family Standard) [7.8.7]. The scope of this standard covers the switching equipment, non-radio transmission equipment (such as multiplexers, SDH, PDH, ATM, DCC systems), power supply equipment and supervisory equipment. (Beside this standard, also other product standards include EMC requirements. For instance, special requirements are applicable for cable television systems.) The requirements according to the standard have been defined in a way so that the equipment has sufficient immunity. There might be cases with low probability, when the level of disturbance is higher than the immunity level specified in this standard. In such cases special mitigation measures shall be made.

According to the installation environment the standard is categorising the equipment into the following two groups:

- Equipment installed in telecommunication centres;
- Equipment installed in locations other than telecommunication centres.

Into the second category belongs, for instance, equipment installed in customer premises and outdoor locations such as street cabinets. In such cases, because of the non-controlled environment, the requirements are more severe.

In the followings – not aiming at the complexity – we introduce what kind of EMC tests are defined in the standard. The individual tests can be given as subsection, according to the ports to be tested:

- Testing of the enclosures (see Table 7.8.1.);

- Testing of ports of outdoor telecommunication signal lines (see Table 7.8.2.);
- Testing of ports of in-house telecommunication signal lines;
- Testing of ports of alternating current power supply;
- Testing of ports of direct current power supply.

Beside immunity and emission, this standard is including also the requirements of „resistibility". (The resistibility means the ability of telecommunications equipment to withstand the effects of certain electrical, magnetic and electromagnetic phenomena without being damaged.) The following performance criteria serve for the control of compliance with the requirements of immunity and resistibility:

- Criterion „A": The system shall continue to operate as intended; usually no degradation of performance or loss of function is allowed;
- Criterion „B": The system shall continue to operate as intended after the test. Usually no degradation of performance or loss of function is allowed. During the test degradation of performance may occur, but no change of actual operating state or stored data is allowed.
- Criterion „C": Temporary loss of function is allowed, provided that the equipment is self-recoverable.
- Criterion „R": The equipment shall withstand the test without damage or other disturbances (such as corruption of software) and it shall operate properly according to the specified parameters after the transient electromagnetic phenomenon has ceased. The test may cause the operation of protection or fuses. After having them replaced or reset, the normal operation shall restore.

	Environmental phenomenon	Test levels and characteristics	Reference	Performance criterion	Remarks
Immunity					
1.	Electrostatic discharge	Contact discharge: 4 kV Air discharge: 4 kV	EN 61000-4-2	B	
2	Radio frequency electromagnetic field amplitude modulated	80 – 1000 MHz 3 V/m 80 % AM (1 kHz)	EN 61000-4-3	A	
Emission					
3.	Radiated Electromagnetic field at 10 m	30–230 MHz: 40 dB (μV/m) 230-1000 MHz: 47 dB (μV/m)	EN 55022	Not applicable	Large systems should be tested according to ETS 300 127
Resistibility					
4.	Electrostatic discharge	Contact discharge: 8 kV Air discharge: 15 kV	EN 61000-4-2	R	

Table 7.8.1. EMC test of equipment installed in telecommunication centresEnclosure port

Ports for outdoor signal lines

	Environmental phenomenon	Test levels and characteristics	Reference	Performance criterion	Remarks
Immunity					
1.	Fast transients	0,5 kV; 5/50 ns waveform 5 kHz rep. frequency	EN 61000-4-4	B	
2.	Surges	1 kV; 10/700 μs waveform	EN 61000-4-5	B	
3.	Radio frequency conducted, continuous	0,15 – 80 MHz; 3 V; 80% AM (1 kHz)	EN 61000-4-6	A	
Emission					Not applicable
Resistibility					
4.	Surges	4 kV; 10/700 μs waveform	ITU-T Recommendation K.20	R	Applies to cables longer than 500 m with primary protection
5.	Surges	1 kV; 10/700 μs waveform	ITU-T Recommendation K.20	R	Applies to cables longer than 500 m
6.	Power induction	300 V; 50 Hz; 200 ms	ITU-T Recommendation K.20	R	Applies to cables longer than 500 m

Table 7.8.2. EMC test of equipment installed in telecommunication centres

("Manual" intervention is allowed.)

Criterion „A" is applicable in case of continuous phenomena, criteria „B" and „C" in case of transient phenomena, and criterion „R" in case of resistibility tests.

Health protection and life safety regulations: Since the beginning of its evolution, the biosphere is subjected to the impacts of its environment. Due to the technical development, a wide spectrum artificial radiation with growing intensity is added to that. The natural RF radiation, which, first of all, is coming from the Sun, reaches the Earth with an intensity of less than 0.01 mW/m^2. Against it, the level of the man-made radio frequency interference (commonly known as "Electro-smog") is varying in a range from several time ten $\mu W/m^2$ that can be measured in households to several ten W/m^2 that can occur in certain workplaces [7.8.8].

In the interest to ensure that electromagnetic fields shall not cause adverse health effects, international and national guidelines, recommendations, regulations and/or standards define the basic restrictions and reference levels for intensity. In

general, these levels are determined with the principle that by evaluating the results of observations and experiments the malfunctioning of the central nervous system is stated and/or the value of the field strength or power density causing undesired overheating on the body surface or in the internal organs, is defined, and than this value is divided by a certain safety factor.

In connection with the above mentioned effects it needs to be mentioned: competent studies dealing with the results of epidemiological investigations emphasise that to date no correct analyses could provide convincing evidence of interrelation of RF exposure and risk of increased incidence of cancer.

In the latest recommendations and specifications - beside or instead of 'field strength' and 'power density' - the term of SAR (**S**pecific energy **A**bsorption **R**ate) averaged over the whole body or over parts of the body, is defined as the rate at which energy is absorbed per unit mass of body tissue and is expressed in watts per kilogram (W/kg).

In Hungary the Standard MSZ 16260-86 issued in the beginning 1987 and modified in 1993, is in force. The limits of the Standard are given in Table 7.8.3.a and Table 7.8.3.b. [7.8.9].

It is worth mentioning that the limits specified in the Hungarian standard are significantly lower than those contained in the EU Recommendation.

For the sake of historical correctness it is worth to mention what happened with the European Prestandard ENV 50166 that has been prepared by CENELEC. In 1994, still as a preliminary standard prENV 50166, it was distributed among CENELEC members with the request to submit their comments. In 1995, it was

Zone	Electric field strength, V/m		
	30 kHz - 3 MHz	3 - 30 MHz	30 - 300 MHz
Non-hazardous	3	3	3
Safety	50	30	20
Work-	120	60	40
Work limited in time	960/t*	480/t*	320/t*
Hazardous	1000	600	400
* Whereas t = duration of time in hours spent within the given zone in one calendar day. The value of t shall not exceed 8 hours.			

Table 7.8.3.a - Electric field strength limits according to Standard MSZ16260-86

published in two parts, as European Prestandard:

ENV 50166-1 Human exposure to electromagnetic fields,

Low frequency (0 Hz to 10 kHz)

ENV 50166-2 Human exposure to electromagnetic fields,

High frequency (10 kHz to 300 GHz)

The Prestandard provides detailed information on the indirect and direct effects of low and high frequency electromagnetic fields, the possible sources, the methods for summary of effects of simultaneous exposure from more sources, the calculation and measurement methods, and of course the limits.

Considering that after its prolongation in 1997, it had not been finalised either in 1999, it has lost its effect. Between 1997 and 1999 it has been endorsed and issued also as a Hungarian National Prestandard. The limit values of it can be used even today as good references.

The Council of Health Ministers of the European Union on its meeting in June 1999 in Luxembourg has dealt with the Recommendation on the limitation of exposure of the general public to non-ionising electromagnetic radiation. The Council's work has drawn on the preparatory works of the International Commission on Non-Ionising Radiation Protection (ICNIRP) [7.8.10], and in connection with the adoption of the Recommendation some basic principles have been fixed [7.8.11]. In line with one of these important principles, the Recommendation leaves it to the Member States the classification of the radiation sources and the limitation of their emission. On 12 July 1999, the Recommendation was approved under Ref. no.: 1999/519/EC by the Council of the European Union on its meeting in Brussels [7.8.12]. The limits, basic restrictions in the Recommendation are specified first of all for the general public, the values for workplaces - in the spirit of the Union's Treaty - shall be determined in the Member States by the political decision making mechanisms.

The way of problem handling and the considerations of the Recommendation reflect the approaches of ENV 50166. Two restrictions are differentiated:

Zone	Power density, mW/cm^2	
	Standing radiator	Rotating, or scanning radiator
Non-hazardous	-	-
Safety	0,01	0,1
Work-	0,1	1
Work limited in time	$(0{,}08 / t^*)^{1/2}$	$(8 / t^*)^{1/2}$
Hazardous	10	100
* Whereas t = duration of time in hours spent within the given zone in one calendar day. The value of t shall not exceed 8 hours.		

Table 7.8.3.b – Power density limits in frequency band 0,3-300 GHz according to Standard MSZ16260-86

The basic restrictions have been calculated so, that 1/50 of the intensity values causing according to the research studies acute (nervous system and/or thermal) effects has been taken into consideration.

The frequency ranges of the basic restrictions are tallying with the following considerations:

- between 0 and 1 Hz basic restrictions are provided for magnetic flux density for static magnetic fields (0 Hz) and current density for time-varying fields up to 1 Hz, in order to prevent effects on the cardiovascular and central nervous system,
- between 1 Hz and 10 MHz the observation of basic restrictions provided for current density makes it possible to ensure normal functions of the nervous system,
- between 100 kHz and 10 GHz basic restrictions on SAR are provided to prevent whole-body heat stress and excessive localised heating of tissues,
- between 10 GHz and 300 GHz basic restrictions on power density are provided to prevent heating in tissue at or near the body surface.

The basic restrictions (limits) are shown in Table 7.8.4.

Reference levels are such, in physical quantities well measurable, limit values, the respect of which ensures that the basic restrictions will not be exceeded, even in the case of closest coupling of the field to the exposed human body. The other important role of the reference levels summarised in Table 7.8.5 and Table 7.8.6 consists in that they give orientation for the determination of the emission limits specified in the different series of standards detailed under section 7.8.2.

It might be of interest to pick out from the wording and make reference on two important statements of the Recommendation:

Frequency	Magnetic flux density, mT	Current density (rms*), mA/m^2	Whole body SAR (average), W/kg	Localised SAR (head, trunk) W/kg	Localised SAR (limbs), W/kg	Power density, (S), W/m^2
0 Hz	40	-	-	-	-	-
0-1 Hz	-	8	-	-	-	-
1-4 Hz	-	8/f**	-	-	-	-
4-1000 Hz	-	2	-	-	-	-
1-100 kHz	-	f/500	-	-	-	-
0.1-10 MHz	-	f/500	0,08	2	4	-
10 MHz-10 GHz	-	-	0,08	2	4	-
10-300 GHz	-	-	-	-	-	10

Table 7.8.4 Basic restrictions for general public according to EU Recommendation

*Averaged over a cross section of 1 cm^2 perpendicular to the current direction

**f = frequency in Hz

***All SAR values are to be averaged over any six-minute period

- Actions on limiting the exposure of the general public to electromagnetic fields should be balanced with the other health, safety and security benefits that devices emitting electromagnetic fields bring to the quality of life, in such areas as telecommunications, energy, life- and public security.
- Member States, in order to enhance understanding of risks and protection against exposure to electromagnetic fields should provide, in an appropriate format, information to the public on the health impact of electromagnetic fields and the measures taken to address them.

The literatures referred to under [7.8.13] and [7.8.14] provide a great help in the comparison of the national standards regulating this field. The considerations related to mobile handsets and base stations are not discussed here in this study, only a reference is made to the best known source of information on this subject matter [7.8.15]. We only draw the attention to the Hungarian National Standard dealing with the radiation safety requirements of laser products [7.8.16].

7.8.5. Tests

The type approval test carried out under laboratory conditions serves for the certification of EMC. The basic standards of series EN 61000-4-X give provisions for testing and measurement techniques. In the frame of these series there have been

Frequency	E-field strength, V/m	H-field strength, A/m	B-field, μT	Equivalent plane wave power density, W/m^2
0-1 Hz	-	3.2×10^4	4×10^4	-
1-8 Hz	10 000	$3.2\times10^4/f^2$	$4\times10^4/f^2$	-
8-25 Hz	10 000	4 000/f	5 000/f	-
0.025-0.8 kHz	250/f	4/f	5/f	-
0.8-3 kHz	250/f	5	6.25	-
3-150 kHz	87	5	6.25	-
0.15-1 MHz	87	0.73/f	0.92/f	-
1-10 MHz	$87/f^{1/2}$	0.73/f	0.92/f	-
10-400 MHz	28	0.073	0.092	2
0.4-2 GHz	$1.375\ f^{1/2}$	$0.0037\ f^{1/2}$	$0.0046\ f^{1/2}$	f/200
2-300 GHz	61	0.16	0.20	10

Table 7.8.5. Reference levels for general public according to EU Recommendation

*Frequencies are expressed in the relevant measuring unit of the frequency range investigated

issued so far standards dealing with the following issues, and these reflect quite well to the great variety of phenomena to be taken into consideration:

1. Overview of tests;
2. ESD immunity tests;
3. Radiated radio-frequency field immunity test;
4. Electrical fast transient / burst immunity test;
5. Surge immunity test;
6. Conducted radio-frequency immunity test;
7. General guide on harmonics and interharmonics measurements for power supply systems and equipment connected thereto;
8. Power frequency magnetic field immunity test;
9. Pulse magnetic field immunity test;
10. Damped oscillatory magnetic field immunity test;
11. Voltage dips, short interruptions and voltage variations immunity tests;
12. Oscillatory waves immunity tests;
14. Voltage fluctuation immunity tests;

Frequency	Maximum contact current, mA
0 Hz-2.5 kHz	0.5
2.5-100 kHz	0.2 f*
100 kHz-110 MHz	20

*kHz

Table 7.8.6 Contact currents for general public according to EU Recommendation

15. Flicker test;

16. Conducted common mode disturbance immunity test in the frequency range 0-150 kHz;

24.Test methods for protection devices for HEMP conducted disturbance;

27. Unbalance, immunity test;

29. Voltage dips, short interruptions and voltage variations on d.c. input power port immunity tests.

(The list is not complete, because this series of standards is under continuous future development .)

In section 7.8.4.1 the tests of wired telecommunications networks have already been discussed. Regarding wireless systems, some standards, elaborated by ETSI, worth mentioning as well, which specify both limits and measurement techniques:

EN 300 279: Conducted and radiated disturbance signals of Private land Mobile Radio equipment

ETS 300 329: Emission and immunity tests of DECT equipment

ETS 300 339: Measurement methods and limits for radio communications and connected ancillary equipment, regarding which a harmonised EMC standard does not give stipulations

ETS 300 340: Emission and immunity tests of ERMES receivers

EN 300 341: Emission of Land mobile radio communications equipment

ETS 300 342: Emission and immunity tests of GSM and Digital Cellular Systems (DCS)

EN 300 385: EMC measurements and limits for Fixed Radio Links

EN 300 390: Emission of radios with built-in antenna of Land mobile radio communications systems

ETS 300 447: Emission and susceptibility of VHF FM broadcasting transmitters

EN 300 673: EMC measurements for terrestrial VSAT equipment

ETS 300 717: EMC tests for public analogue cellular radio communications equipment. Mobile and portable equipment

ETS 300 826: 2.4 GHz wide-band transmission systems and HYPERLAN

EN 300 827: EMC measurements and limits for TETRA equipment

7.8.6. Guidelines aiming at ensuring electromagnetic compatibility

There are many types of installations and electromagnetic compatibility can be successfully achieved through different approaches. The IEC 61000-5-1 standard [7.8.17] recommends a general approach, by the application of which special mitigation methods might not be necessary when the equipment satisfies applicable emission and immunity requirements.

The process adopted for ensuring electromagnetic compatibility of equipment and installations may take two approaches, depending on how early in the design, and implementation the EMC requirements are taken into consideration.

- At the early stages of major installations, each compatibility level (specific for a given electromagnetic disturbance) can be assigned for the particular environment of the installation or equipment. Through the application of the overall mitigation schemes, the apparatus and its installation practice are then specified with immunity and emission levels corresponding to the predetermined compatibility level.
- At later stages of the design, for the installation of additional apparatus or the initial installation of commercially available apparatus, for which no opportunity exists to modify its EMC characteristics, a mismatch may occur between the overall, *de facto* compatibility level of the site and the capability of the apparatus. In such a case, mitigation methods shall be selected, in order to close the gap between the environment and the apparatus immunity levels to a minimum.

Electromagnetic disturbances are caused by conducted or radiated phenomena. An apparatus can be both the emitter and the susceptor (potential „victim”) at the same time.

With regard to electromagnetic compatibility there are three main areas that can be considered:

- emitters: sources of disturbances, influenced by the design of the apparatus;
- coupling paths: influenced by the installation practices;
- susceptors: influenced by the design of the apparatus.

In order to assure EMC, three types of measures should be applied, as necessary:

- at the emitter: reduction of emission;
- at the coupling: reduction of coupling;

- at the susceptor: increase of immunity.

In the interest to provide a transition from the overall concept by analysis the interrelation between the environment and the apparatus and to come to the application protection method, the ports of the apparatus (which are the interface of the apparatus toward the external electromagnetic environment) shall be investigated. The various EM disturbances enter or exit the apparatus through these ports. By identifying such ports, protective steps can be specifically related to the nature of the EM phenomenon, its coupling path and its impact on the functional elements of the apparatus (immunity) or on the environment (emission).

Figure 7.8.2 introduces the input ports of electromagnetic disturbances.

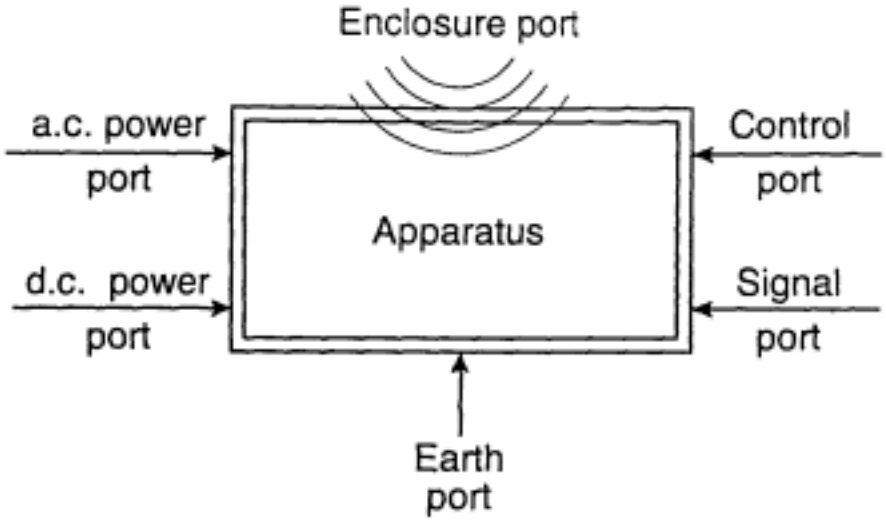
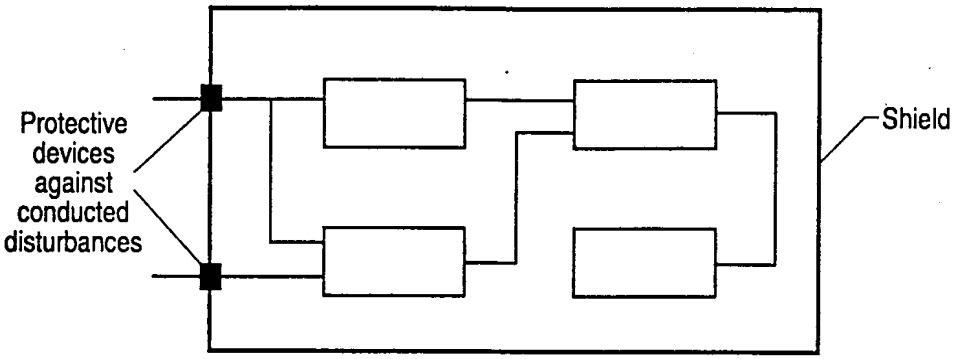


Figure 7.8.2: Representation of equipment ports interfacing with the electromagnetic environment

For the electromagnetic compatibility appropriate mitigation steps should be applied on every port of the apparatus (system, installation).

There are two general approaches to obtain EMC immunity for an installation, either by a global protection (figures 7.8.3. and 7.8.4.) or by a distributed protection (Figure 7.8.5). In certain cases, mitigation methods might not be necessary, for instance, if the equipment has a sufficiently high immunity level, compared with the prevailing disturbance level.



igure 7.8.3: Global protection by single barrier

NOTE: According to the principle of a single barrier, mains filters, surge-protective devices and a shield protect the whole installation. No specific protection is applied to the individual units, except when internally generated disturbances exist.

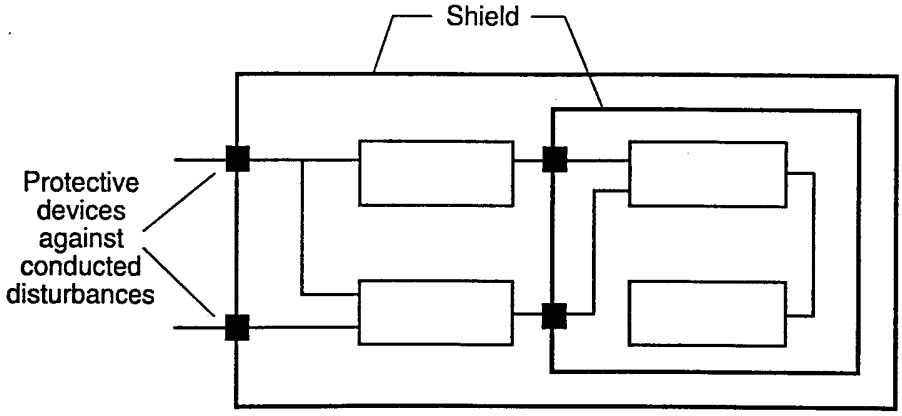


Figure 7.8.4: Global protection by multiple barriers

NOTE: According to the principle of multiple barriers, no specific protection is applied to the individual units, but there is a cascading of multiple electromagnetic barriers according to the susceptibility level(s) of the units.

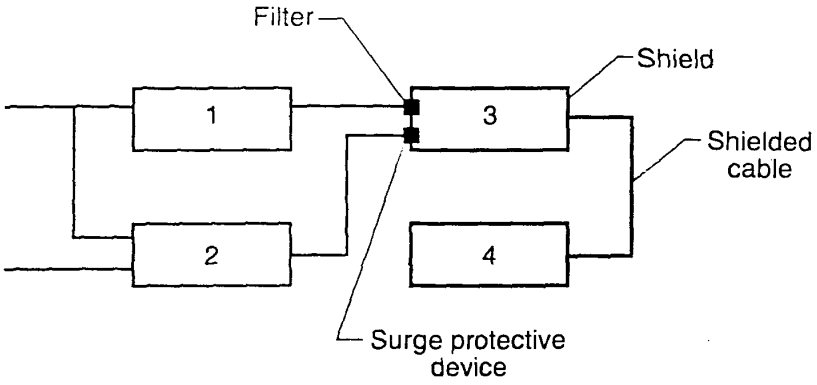


Figure 7.8.5: Principle of distributed protection

NOTE: According to the principle of distributed protection, units 1 and 2 are not protected, only units 3 and 4, which contain sensitive electronics, are protected. For the protection of the latter ones, specific enclosures, filters, or other protective devices and shielded cables are used.

List of abbreviations

CENELEC European Electrotechnical Standardization Committee

CISPR International Special Committee on Radio Interference

EMC Electromagnetic Compatibility

EMI Electromagnetic Interference

ERM Electromagnetic Compatibility and Radio spectrum Matters (ETSI Technical Committee)

ESD Electrostatic Discharge

ETSI European Telecommunications Standards Institute

HEMP High Amplitude Electromagnetic Pulse

ICNIRP (**I**nternational **C**ommission on **N**on-**I**onising **R**adiation **P**rotection)

IEC International Electrotechnical Commission

ITU International Telecommunication Union

LEMP Lightning Electromagnetic Pulse

NEMP Nuclear Electromagnetic Impulse

SAR Specific energy Absorption Rate

Bibliography

[7.8.1] MSZ IEC 50(161) International Electrotechnical Vocabulary (IEV), Chapter 161: Electromagnetic compatibility (1994)

[7.8.2] MSZ IEC 1000-1-1: Electromagnetic compatibility (EMC), Part 1: General, Section 1: Application and interpretation of fundamental definitions and terms

[7.8.3] Directive 89/336/EEC (Electromagnetic compatibility), (1989)

[7.8.4] Guide to the application of Directive 89/336/EEC, (1997)

[7.8.5] Joint Decree No. 31/1999. (VI.11.) GM-KHVM on Electromagnetic Compatibility

[7.8.6] CENELEC Report: R210-001 EMC standardisation for product committees (January, 2000)

[7.8.7] EN 300 386-2: Electromagnetic compatibility and Radio spectrum Matters (ERM); Telecommunication network equipment; Electromagnetic Compatibility (EMC) requirements; Part 2: Product family standard (1997-12)

[7.8.8] WHO: Electromagnetic Fields and Public Health, Fact Sheet N183, http://www.who.int/inf-fs/en/fact183.html

[7.8.9] MSZ 16260-86, Permissible limits of high frequency electromagnetic fields

[7.8.10] International Commission on Non-Ionising Radiation Protection: Guidelines for limiting exposure to time-varying electric, magnetic, and electromagnetic fields (up to 300 GHz), Preprint scheduled to appear in Health Physics April 1998, Volume 74, Number 4, pp. 494-522

[7.8.11] European Union, Council of Health Ministers: Adoption of a recommendation on the limitation of exposure of general public to electromagnetic fields
http://europa.eu.int/comm/health/ph/news/old/electro_en.htm

[7.8.12] COUNCIL RECOMMENDATION of 12 July 1999 on the limitation of exposure of the general public to electromagnetic fields (0 Hz to 300 GHz), (1999/519/EC) Official Journal of the European Communities, 30.7.1999

[7.8.13] Mátay G., Zombory L.: Physiological impacts of radio frequency radiation and its applications in medicine and biology, University Learning Book
Publishing House Budapest Technical University, Budapest, 2000

[7.8.14] Thuróczy Gy.: Radiation health aspects of mobile communications, 'Magyar Távközlés' July 1998, Volume IX., Number 7, pp. 26-33.

[7.8.15] International Commission on Non-Ionising Radiation Protection: Health Issues Related to the Use of Hand-Held Radiotelephones and Base Transmitters, Health Physics April 1994, Volume 70, Number 4 pp. 587-593

[7.8.16] MSZ 16261, Radiation safety requirements of laser products

[7.8.17] MSZ IEC 61000-5-1: Electromagnetic compatibility (EMC) Part 5: Installation and mitigation guidelines, Section 1: General considerations. Basic EMC publication

[7.8.18] IEC TS 61000-1-2 : General – Methodology for the achievement of the functional safety of electrical and electronic equipment with regard to electromagnetic phenomena

7.9. Equipment and system selection considerations

Gábor Nagygyörgy dr., author

Kántor Csaba dr., reviewer

7.9.1. Introduction

The purpose of the present chapter is to give a short overview upon the major technical criteria of system selection, but some other considerations, which – in addition to the technical requirements – influence our choice, shall be highlighted as well. Among them, the reliability of the vendor/supplier from market, or financial points of view, or the life-time costs, etc. can be mentioned.

At making a choice of a telecommunication system, the first task of the planner/designer is to determine and specify the main functions of the system, i.e. to decide what is needed. This usually includes a technical analysis based on marketing survey, a system planning. On the basis of the large-scale and later the detailed system plan, the buyer issues a tender invitation. By the assessment and evaluation of the bids received the most appropriate telecommunication system can be chosen on the basis of the analysis of the technical and financial criteria. The system selection means, therefore, such a technical/financial analysing and planning process, in the course of which the buyer – with regard to the market conditions – brings into harmony his needs with the possibilities. In the following chapter we give an overview upon the major steps of this process and the key criteria of (system) selection. Let’s clarify first, what is meant by a system.

System (A): a system shall mean the totality of objects being in interrelation with each other, including their interrelations, structure and dynamics.

Telecommunication system (B): a telecommunication system shall mean an organised totality of a certain set of terminal equipment, switching and transmission technical devices for making a service provisioning possible.

7.9.2. Technical aspects of system selection

The planning process described below is for the case of development of an absolutely new system. In the practice, usually new partial, sub-systems are added, adapted to an already existing system, or existing systems are expanded or extended. If it is the case, then certain steps are reasonable omitted from the hereunder described procedures of specification and selection.

7.9.2.1. Large-scale system plan

At system planning the parameters – technical requirements – of the telecommunications system in question are determined with a “from top to bottom” approach. It means that first we are modelling from endpoint to endpoint the whole info-communication connection (in a “high-level model”, “large-scale macro-model”, “large-scale system plan”), on the basis of which – “going downwards” – we determine and allocate the individual system components, elements and the requirements of equipment.

In the large-scale system plan, there shall be specified the signaling, the signal levels, circuit parameters from subscriber to subscriber. In the macro model, all of terminal equipment and network resources constitute a customer-to-customer integral whole, whereas the individual system elements possibly do not have a requirement of their own. The large-scale (high-level) particularities are often prescribed in ITU-T, ETSI or other standardised systems, reference models, but making of our own, non-standardised, individual models is also allowed. Reference network requirements are included also in the fundamental technical plans issued by the national regulatory body (HÍF). The standard reference models sometimes include recommendations for the allocation of parameters of the system elements, as well. The planner has to design the actual system in accordance with the guidelines given by the reference model. The large-scale planning is thoroughly discussed in other chapters of the book. Creation of models is a very time-consuming task, the standardisation bodies are working on their appropriate reference models over years. The planning of non-standardised systems, therefore, requires lot of time and increases the risk of the telecommunication service providers (see later).

The most important aspect of the selection is the implementation of the operation (functions) provided by the system. The planner has to quantify the needs of the customer, transform them into the terms of technical requirements. Such customer needs are the volume of information, the speed of access to the data, the soundness of the data, the usability, reliability of the connection, the expected volume of traffic, the comfort at use, etc. In addition to it, further functional requirements are determined by the financial, economical interests of the telecommunication provider, or by the legislative or authority regulations and stipulations. Here we can mention: the cheap and well manageable operability and maintenance, the billing possibilities, the customer's safety and security, or the noise load on the environment, etc. With due analysis of the functional demands – the logical structure (architecture) of the system can be built up with regard to the different aspects of planning and to the requirements of standards and norms, and in this way the technical requirements for the whole complex system can be specified. In a great number of cases the planner/designer can find a standardised model (reference) system which is suitable for the satisfaction of the functional requirements.

A large-scale system plan is modelling the connection from endpoint to endpoint, the requirements of it describe the technical parameters of the whole connection. The groups of the major requirements of system plans are usually specified in standards. The main categories of the requirements are:

- *Structure of the system, i.e. the logical and functional system architecture.* It includes the system building blocks, the system components as well, along with the location and relations of the sub-systems and equipment. A telecommunication company needs not only the telecommunication functions, but also the support systems which support the servicing of the customers and operation of the network. A state-of-the-art telecommunication system today is planned, designed, constructed and purchased only along with network management, O&M support and billing systems. We have to give the supplier the large-scale system plan (reference model, architecture) for the systems we are intended to purchase, on the basis of which we have specified the requirements for the system and the individual system elements.
- *Quality of the transmitted signals*: signal levels, signal bit rate, signal shape, signal distortion and modulation noise, reflection, attenuation, etc.
- *Continuity of the connection*: susceptibility, drop-out time, time to failures.
- *Message routing, message handling*: signals and protocols, synchronisation.

- *Network management*: traffic routing, monitoring, alarms and security signals, etc.
- *Serveability performance* (number of subscribers, traffic capacity,) and expandability, trafficability and overloadability.
- *General requirements of sub-systems*, like, for instance, the general requirements of NMS, tarification and billing/accounting systems linked to the telecommunication system.
- *Security system requirements*: control and limitation of access, security system administration, etc.
- *Interface requirements* (integration, adaptation requirements) towards the existing systems, or other platforms, or towards the network management and billing systems,

and in addition any other – above not mentioned – technical requirement, that in general the implemented network and platform shall meet in the interest of establishment and maintenance of a continuous telecommunication connection and proper servicing of customers.

Since the telecommunication systems usually consist of equipment accommodated in different environment and of transmitting media among them, it is possible only rarely to fix requirements also for the installation conditions of the complex system in the large-scale plan. The installation, implementation requirements of the general system plan are applicable for the first implementation and for the deployment. A well-designed telecommunication system can be expanded, extended gradually, in small steps, according to the needs, even by subscriber or by channel. Possible the system shall be so construed that also the first implemented segments can provide full functionality.

7.9.2.2. Detailed system plan: system element-related requirements

The detailed system plan can be produced – making due planning considerations – from the general network requirements. In the course of making the detailed planning, the requirements of the sub-systems, system elements, modules can be defined, allocated. Similarly, first the general functional requirements of the system elements and modules are defined in a way, so that they match to the operation of the whole system. The system elements are the elements of the physical network, and at the same time they are telecommunication equipment. Therefore in the course of the detailed system planning, we have to determine expectations,

requirements also for the physical design, outlook, construction, as well as for the installation of the system elements. The system elements may be hardware and software elements or their combinations.

The system element requirements are defined according to the following categorisation:

- Allocated transmission or signal processing requirements (e.g.: insertion loss, allowed drop-out, noise, etc.),
- Individual operational requirements, (e.g.: signal processing, disturbance/noise emission, etc.)
- Software requirements and
- Equipment-related requirements, as to their:
- Construction,
- Accommodation,
- Operation and Maintenance
- Installation-related design, transportation, mounting and environmental requirements.

With an object oriented approach, it can be stated that the software programs, having their own characteristics, features, are similar, equivalent system elements of a telecommunication system, like the equipment. The software selection and the specification of the software requirements play a very important role from the point of view of the operation and quality of the whole info-communication system. The software elements of the telecommunication system and the requirements of them shall be specified on the basis of the large scale and the detailed system plans as much detailed as it is already usual at the specification of the equipment's construction requirements. Reference [C] gives a quite rich source of literature information and of standards covering this subject matter.

The software programs of telecommunication equipment are very often based on standards (for instance the protocols), in such cases, instead of a detailed system requirement, it is sufficient to refer to the given standard only.

General software requirements are: portability, easy usage, re-usability, accuracy, testing, maintainability, possibility of modification (corrective-, adaptive- and improvement modifications), reliability, efficiency, integrity and finally the deepness, comprehensiveness of documentation.

When making the software specifications, especially the followings shall be taken into consideration:

- The platform on which the software should work.
- Other software systems, with which the software shall interwork, i.e. what are the software interface requirements.
- Remote software monitoring and remote software downloading shall be specified, as a requirement.
- Reasonably the software should be intelligent enough to be able to recognise, identify and possibly prevent the false controlling commands. According to certain statistics, 40% of outages of the modern, sophisticated telecommunication networks is attributable and originated from the issue of false, mistaken commands.
- The software reliability, or the required conditions of reliability assurance.
- Safety of the software
- A special attention shall be given to the software security. The software shall have such a protection system, which recognises the attacks, registers them, or probably can eliminate them. Nowadays, one of the most risky problems in the operation are the malicious attacks against the telecommunication systems. The successful attacks cause harm of reputation of the firm but on a longer time they may lead to that that the subscribers, customers go to another service provider.
- The existence of a software supporting organisation shall be required. This often means a multi-level organisation, which shall render a 365-day, 8760-hour, i.e. continuous (round-o-clock) assistance to the operator of the system.
- Training of the software operators shall be required, as well.

7.9.2.3. Operational requirements

The telecommunication network shall have a *network management sub-system*. The O&M system shall make the “remote alarms”, remote diagnostics, fault detection and fault localisation possible, as well as the re-routing of the traffic from the defective segments to alternative ones. In addition, the fault correction activities shall be supported by fault correction task issuing and feedback, reporting sub-systems (e.g.: workflow programs). There shall be taken into consideration the impact of timescaling of the expansion of network operation and the telecommunication system onto the building up of the management sub-system. For the remote supervision, monitoring of the telecommunication system it is necessary that each and every system element is adapted to the network management sub-system. The type of the management system interface shall be specified. Also the

equipment shall support the remote diagnostics, remote fault detection and regular operational testing (without causing an observable interruption in the operation).

The *availability* of the services is influenced by the reliability of the network. A service network with acceptable availability can be built-up even from less reliable system elements, if the network has hot reserves, alternative routes and paths. Chapter 7.1 of our book discusses in details the interrelations between the reliability of the system elements and that of the network and the relevant planning aspects.

The permitted outage parameter values for the system elements and equipment can be usually calculated, or derived and allocated from the system structure and from the availability requirements of the services. The reliability of the equipment has financial/economical aspects as well, since the more reliable equipment is often more expensive. The reliability allocated to the individual elements is influenced also by the useful life-time costs, or operational considerations: the operation of a more reliable equipment is cheaper, since the probability of outage of such a unit is lower. The system designer has the possibility to optimise the total costs of the whole system with reasonable modification of the system structure, or topology (alternative paths, routes, reserves, etc.).

In case of development of a green-field system, a Billing Sub-System plan shall be made, as well. The details of it – reasonably – are almost the same like the telecommunication system selected according to the above described criteria.

In case of an already existing, known billing system, we prescribe that the new telecommunication system shall have appropriate interfaces towards the tarification and billing sub-systems.

7.9.3. Design and construction requirements of equipment

7.9.3.1. Construction

The construction of system elements shall be designed with consideration to the criteria of installation. The installation into telecommunication building can be judged first of all on the basis of financial and operational safety considerations. The ease of mounting of the device is one of the important parameters of installation. The modular construction, the easy interchangeability of the components, the

arrangement of and accessibility to the moduls, their stability and protection against tilting over, support and facilitate the operability and maintenance. The possibilities for the accommodation of cables shall be specified both for the telecommunication buildings and for the customer premises. The design and construction of the cooling block, ventilation is important, similarly to its adjustment to the building's air-conditioning system. In addition, the network elements accommodated in the customer's premises shall comply with severe security, safety and esthetical requirements, too.

The investment demand of a telecommunication building installation can be reduced by the reduction of the size of the equipment, by the application of an appropriate form (with proportionate dimensions of the height, width and depth), and low weight- and floor load. At the same time we have to consider that the specific dissipated heat in space of an equipment of smaller size may be higher, which may require more intensive cooling.

The network termination equipment accommodated at the customer's premises shall be as small as possible and protected against environmental impacts.

In telecommunications building installations the equipment are well protected. In outdoor, public places, however, strict protection requirements shall be enforced against the enclosure of the equipment: it shall provide protection against impact, drip, dust, insects, unauthorised access/intrusion, breaking up and line-tapping.

Environmental resistance is standardized in every countries.

Any system element shall operate normally according to its specifications at the place of its installation. The criterion of the selection is that the equipment shall withstand the ambient, environmental operation conditions both indoor in temperature controlled or non-temperature controlled locations and also in outdoor and underground locations in case of stationary or mobile accommodation. The environmental resistance requirements shall be specified for transportation and storage as well. The easiest way to specify the environmental resistance is, making a reference to the relevant standards. The ETS 300 019 multi-part standard series of ETSI, reducing the range of standards specified in the well-known IEC 723 standard series, include a set of environmental requirements applicable for the European climatic conditions. The standards provide recommendations for the requirements of

ambient temperature and air conditions (solar radiation, operation in a heat-trap, etc.), and that of mechanical, climatic, atmospheric, biological, chemical hazards and impacts (like groundwater, gases and impurities, etc.).

Environmental protection criteria is defined in accordance with the possible accomodation

Nowadays it is a very strong expectation from the side of the society that the equipment, facilities of a telecommunication system, if they are accommodated outdoor, shall fit into the landscape. Therefore, such equipment shall be installed in places, where they cannot be seen, or if it is not possible, for e.g. in case of radio equipment, an equipment having a camouflaged enclosure, shall be selected. It is an important criterion of the selection, that quantity of contaminating waste materials during the installation and later the operation shall be minimised. The system elements must not include any radioactive substance (even not in ionizators) and the mounting, installation waste materials shall not be poisoning, either. Similarly, the high frequency signal emission (radiation hazard) might be also an important environmental factor.

Electromagnetic compatibility (EMC) requirements are set forth in standards and legal rules. In Hungary the equipment shall comply with the disturbance emission standards specified in the Decree of the Minister of Economy. In accordance with the standards also the (electromagnetic) susceptibility of the system elements shall be specified, namely according to the requirements of the installation environment. Electromagnetic disturbances may intrude or be emitted through the telecommunication and the energy supply interfaces as well, therefore the EMC requirements shall be enforced against all such interfaces. The key criteria of the selection are: susceptibility (radiated, conducted /low and high frequency), emission (radiated, conducted / low and high frequency, modulated), protection against discharges, fast transients (bursts) and surges.

7.9.3.2. Energy supply and consumption

The energy is becoming more and more expensive, the daily usage time of informatics systems is longer and longer, because of the spreading of on-line services. Due to these, keeping the energy consumption at a low level, is an aspect

at the selection, which is getting more and more important. The energy supply demand of the system element shall be adjusted to the energy supply capabilities of the existing network, in the telecommunication buildings, exchanges, in the access segments (DC power supply, remote feeding), and in the customers' premises, as well (low voltage AC supply). The requirements of the power interface are specified in the knowledge of the existing energy supply system. These requirements shall cover:

Power (feeding) voltage values, operation in normal and abnormal (non-standard) voltage ranges, voltage controll, power consumption. In addition, the protection against the mentioned EMC disturbances (noise emission, protection against noise), as well as protection against transient phenomena (e.g.: protection against lightning), possibility of remote diagnostics, protection against electric shock, shall also be specified.

In this respect special attention shall be paid to the uninterruptable power supply of system elements installed at the customer premises (NTU, terminal equipment). The terminal equipment can be powered either from the mains supply network of the public utility company (Electric Works), or locally. In the second case the required uninterruptable energy supply shall be specified on the basis of the customer's needs and in accordance with the prevailing statutory regulations. It shall be considered as well, whether the costs of energy burden the customer, or the telecom service provider.

Grounding is one of the important safety requirements.

The equipment installed in telecommunication buildings shall comply with the requirements of the national grounding standard in force. In the customer premises, living houses of old construction there are often obsolete grounding solutions applied and various grounding systems are in operation in the different regions of the country. It is reasonable, therefore, to require the availability of the equipment operating in the customer premises in a version, which is adaptable to several grounding systems, so that the service provider has not to re-construct, re-build the grounding of the subscriber's building.

Cooling

The energy dissipated by the equipment may lead to such a temperature rise that endangers the normal, standard operation. Therefore, construction of a reliable cooling system is an important criterion, which influences the availability. The energy consumption of the cooling system is considerable, therefore consideration shall be given to the economical construction of the cooling system, too. If the conditions of installation, accommodation are known, it shall be required that the cooling system of the equipment is adapted to the cooling system of the whole building.

7.9.3.3. Safety requirements

The safety characteristics – serving for the protection of life, health and against loss or damage of properties – of system elements installed in telecommunication exchange buildings, or especially in the customer premises or dwelling units, are prescribed generally in mandatorily observable standards. Such ones are:

Electrical safety requirements, e.g.: protection against electric shock

Surface temperature of covers

Shape of covers: it shall not cause a mechanical injury

Requirements in connection with fire protection:

Resistance to fire and

inflammability of the equipment (i.e. its ability to cause fire);

Existence of fire barriers among the components of a large-sized equipment.

Hazard of radiated energy and shielding of the radiation. It is a requirement that the laser shall not go into the eye, but it shall be shielded against other high frequency radiations as well.

7.9.3.4. Transportation and packaging requirements

It shall be required from the manufacturer of the system that the dimensions, the weight, the design of the equipment shall make it as easy as possible the transportation of the unit to the place of destination and putting it into its final place of

installation. The „transportation resistance” as resistivity against damage in transportation we have mentioned already among the environmental resistance criteria. From the packaging it is expected that the packaging material shall not be harmful to the environment and it shall properly protect the equipment, spare part, etc. from the environmental impacts. The packaging of sensitive spare parts or units shall provide protection against electromagnetic disturbances as well. The packaging shall include the identification labels.

7.9.4. Other considerations

Operations and maintenance support

The operation of a system can be effected either by the owner of it, or by the vendor/supplier or other contractor/partner. The vendor/supplier may help this work by rendering professional advisory assistance, shipping of spare parts, supplementary materials, devices or equipment, operating a repair shop, storage facilities, remote supervision or a system support centre. Reasonably from the technically acceptable systems that one should be chosen, the supplier of which provides the most favourable services, possibly for the entire useful life-cycle of the system delivered. The buyer may determine the intended useful life-cycle of operation and fix it in the contract concluded with the vendor.

Training requirements

The operation of the telecommunication, IT systems is becoming more and more difficult and complicated. Attention shall be paid at the selection that the supplier/vendor provides the necessary training for the operating/maintenance staff, even supplying also the customers with adequate informative brochures, descriptions. Mention must be made also of requiring proper documentation on the system elements.

Qualification of the manufacturer, vendor/supplier

According to law, the buyer is responsible for the quality of the system purchased by him. This responsibility cannot be shifted over to the manufacturer, or vendor/supplier. It means that the telecommunication systems are built-up of

equipment, software programs and other system elements, for the quality of which the telecommunication service provider (buyer, owner of the network) is responsible. The quality of the purchased goods is assured by the purchase quality systems , which systems imply also the qualification/certification of goods and the manufacturers as well. This topic is not the subject of the present book, but those who are interested in deeper study of this issue, may refer to the series of ISO 9000 standards and to the rich literature (for instance: D).

7.9.5. Risks of system selection

In the followings, without aiming at complexity, we call the attention only to some important risk elements. In the interest of minimising the risks, the Buyer – after a comprehensive risk analysis – may set forth guidelines or principles and rules for system selection.

If the planner/designer does not have yet matured ideas, since he/she is looking for a very innovative system, for which at the given time a standardised solution is not yet available, then this increases the risk of the selection. It may happen that the technical evolution takes a different turn in the future, and the selected system becomes very soon obsolete or out of date. The operation, expansion, supplement of non-standardised systems cost always more than that of the system based on commonly used standards.

In the most cases, the system selection means the selection of the manufacturer or supplier/vendor, as well, which results in that the buyer is bound for a longer time to the manufacturer and his systems. Therefore, the right selection of the supplying partner implies a high financial risk. The operation, maintenance, repair of the selected system greatly depend on the supply of reserves and spare parts, on the customer service of the supplier. Similarly, the necessity of later technical enhancement, further development or ecpansion of the system shall be kept in mind, as well. In case of selecting not the right supplier, the manufacturer’s technological backwardness, or if the manufacturer’s company is wound up, the additional development done by a different company, may lead to huge incremental costs. So thus, whether the manufacturer is a well-qualified, financially strong company, or not, is also an aspect of system selection.

Specification of the requirements not in the right way, bears high risk for the customers. The requirements may be inadequate in different ways: Either we specify insufficient requirements, or to the contrary, we prescribe more parameters than it is necessary; at the selection of the system we enforce less strict, or much more strict than necessary requirements. Especially the appropriate specification/selection of the software elements of the system is important, because more an more portion of operational failures of even sophisticated systems is attributable to system failures. In case of purchasing of an underspecified system, it may happen that we do not buy such functions, features that later become important, or to the contrary, maybe that we purchase unnecessary system intelligence, paying extra price and cost for it, but which we cannot utilise. The specification risk can be minimised with value analysis.

If too much weight is given to the purchasing price among the criteria of the selection, it may happen that we will choose a cheap, but from technical points of view not the best system, which will not fill its function and/or will not fit properly to other, already existing systems.

7.9.6. Tendering

The law sets forth a tendering procedure for the selection of elements of a public telecommunication system and for the establishment of such networks. The infrastructure of a given service shall be selected after due consideration of technical and financial conditions. Within the invitation for tenders the Buyer specifies his financial/economical and above described technical requirements.

Evaluation of tenders

The Act on Public Procurements offers enough flexibility for the constructor of the network, who can make his choice not only and exclusively on the basis of the price of the system. The issuer of the tender invitation is not bound by the law to purchase such systems which are obsolete, or cannot be adapted to his existing systems, even not in the case, when the price of them is the best. There are various applicable methods of evaluation to avoid choosing of a technically unsuitable system. According to the one of the well-known procedures, there are two rounds for bidding organised. In the first round the buyers stipulate their technical/economical/financial criteria in connection with the system. Then the

technically acceptable offers are selected and price offers are asked from the best vendors. In the second round the price offers of these vendors are evaluated.

Life-cycle costs

In the most commonly used procedure, the evaluation of the technical/financial and (budgetary) price offers is taking place in one round. It is very important to establish a reasonable weighting proportion between the price and the technical capabilities of the system. At the consideration of the economical/financial aspects, reasonably not the purchasing prices shall be compared, but the life-time costs which include or take into account the purchasing price and the maintenance, repair costs, expenditures that have to be spent on the system over the whole useful life-cycle. Though the life-time costs to a great extent depend on the organisation, efficiency of the provider (buyer), but it is obvious that the network operator is doing this comparison on the same financial/economical basis. (It is easy to see that in case of a system, which requires less maintenance, also the life-time costs of it will be less.)

References

[7.9.1] Információrendszer fejlesztés (Information system development) (Raffai Mária, 1999)

[7.9.2] Távközlő Rendszerek megbízhatósága. Szótár, értelmező szótár (Reliability of telecommunication systems, Explanatory vocabulary.) Hiradástechnikai tudományos egyesület 1974

[7.9.3] Magee, Tripp: Guide to software engineering standards and specifications (Artech House, Boston 1997)

[7.9.4] Ince, D.: ISO 9001 and Software Quality Assurance, McGraw-Hill, 1994

8. REGULATION

In the last decades the pervasive evolution of the telecommunications transformed not only the technologies, networks, services and applications, but also the structure of the players of the telecommunication sector. Twenty years ago the sector was characterized by state-owned national companies in monopolistic situation, which were also administrations. Recently we are witness to the deployment of a competitive environment, the withdrawal of the state ownership, the emergence of newer service providers, transnational service providers and the service providers' alliances, and the continuous rearrangement of their ownership structure. The formation of a multi player market was forced by the technological development. Only in a liberalised competitive market environment, the research and development results could have been really deployed, the multi-coloured variety of the services and applications and the participation of the private capital, thereby the better provision of the telecommunications users could have been produced and graded up.

At the same time, there are limitations of the competition, particularly in oligopolistic situation. The market mechanisms are not able to handle certain aspects of the users expectations, or more generally the public interests in an appropriate way. The most important and characteristic requirements are the followings:

- the advantages expected from the competition are to be presented in reality, i.e. newcomers may enter the market, incumbent operators do not abuse their significant market power, prices should take shape as if perfect competition works;
- the set of essential telecommunication services is to be reached at an affordable price by anybody (universal services);
- the separate telecommunication networks are to be interconnected, the telecommunication services are to be interworking, the quality and security requirements for the services and applications should be performed;
- the scarce resources of telecommunications, such the radio-frequencies required to wireless connections, or the numbers, addresses to routing on the network should be used efficiently and in coordination.

In telecommunications, economic efficiency provided by competition is necessarily supplemented by the enforcement of the public interest, the best methods of that are investigated by expert groups, international organisations, etc.

[8.0.1], [8.0.2], [8.0.3], [8.0.4]. The task of the state is to find an appropriate balance between the economic and public interests, by means of generic and sector-specific acts. Ultimately, legislation, i.e. acts and decrees reflects the policy intention, regulation exerts the professional scope afforded by the legal framework. Legislation and regulation concerning telecommunications in a country have great influence on the evolution of the telecommunication market and services.

Liberalisation, legal establishment of the competition on the telecommunication market involves the creation of a specific, professional-oriented regulatory authority, independent of the service providers. The task of this regulatory authority is to promote the market entry, supervise fairness of competition, enforce the aspects of public interests, and encourage a user-friendly development. The first organisations of this type were established more than half century ago (e.g. Federal Communications Commission, FCC in USA), recently there are more than 100 national regulatory authorities around the world. The regulators of the member countries of the European Community are working within the frame of common directives, e.g. [8.0.5],[8.0.6],[8.0.7],[8.0.8],[8.0.9]. Hungarian telecommunications legislation is conform to these directives [8.0.10].

Recently, evolution leads to the more and more complete convergence of telecommunications, information technology and media sectors, to the formation of a single sector with uniform technology, integrated market promoting the creation of novel synergic services and transformed, allied or just merged organisations. This convergent sector is relevant shaper of the information society and needs the convergence of the regulations of the converging sectors [8.0.7], [8.0.8], [8.0.9]. New forms of the regulation are also emerged, such as self-regulation of the market-players and co-regulation of the regulatory authority and the market-players.

The Chapter first of all presents the three historical models of the telecommunications regulation (periods of natural monopoly, transition to a competitive environment, deployment of the convergence) and the related regulatory tasks and forms. Then the market regulation is discussed, involving the regulatory methods for influencing the conduct of the market players, e.g. the supervision of the market entry and leave, the handling of significant market power to avoid abuse, the rules for setting prices and some issues of universal services provision.

Corresponding to the character of this book, the Chapter goes into more details about system regulation. Technical regulation ensures the network integrity and the interworking of the services. The regulatory issues on the network interconnection and access, as well as the unbundling of local loop are explained. As for the management of scarce resources, the basic rules to the allocation, assignment and usage of identifiers (numbers, names and addresses), as well as radio-frequency spectrum are presented. Finally we cast a glance into the security regulation through the electronic signature for the infocommunication applications.

Gyula Sallai, DSc, Editor of the Chapter

References

[8.0.1] *Melody, William H.* (Ed.): Telecom Reform. Principles, Policies and Regulatory Practices. p.557. Technical University of Denmark , Lyngby, 1997.

[8.0.2] *Intven, Hank* (Ed.): Telecommunications Regulation Handbook. The World Bank (InfoDev Program) és McCarthy Tétrault, Washington DC, November 2000.

[8.0.3] World Trade Organization: Reference Paper on Regulatory Principles, attached to the General Agreement on Trade in Services, Geneva 1997

[8.0.4] International Telecommunication Union: Trends in Telecommunication Reform 1999: Convergence and Regulation. Geneva, 1999.

[8.0.5] European Parliament and the Council: Directive on Radio Equipment and Telecommunications Terminal Equipment and the Mutual Recognition of their Conformity. 1999/5/EC, Brussels, March 1999

[8.0.6] European Commission: Proposal for a Decision of the European Parliament and of the Council on a Regulatory Framework for Radio Spectrum Policy in the European Community (2000) 407, Brussels, 2000

[8.0.7] European Commission: Green Paper on the Convergence of the Telecommunications, Media and Information Technology Sectors, and its Implications for Regulation. Towards an Information Society Approach. COM(97)623. Brussels, 3 December 1997

[8.0.8] European Parliament and the Council: Directive on a Community Framework for Electronic Signatures. COM(1999)93. Brussels, 17 January 2000.

[8.0.9] European Commision: Proposal for a Common Regulatory Framework for Electronic Communications Networks and Services. COM (2001) 380. Brussels, 4th July 2001

[8.0.10] Hungarian Parlament: Act XL/2001 on Communications. Adopted: June 12, 2001, generally effected from December 23, 2001.

8.1. Regulatory Models

Iván Schmideg, PhD, author

Gyula Sallai, DSc, reviewer

Telecommunications regulation is understood as the influence on telecommunications market by legal and economic means in order to reach the goals of the government esteemed important. The government's objectives are determined by the economic, societal, cultural and geopolitical situation, the level of development, and the political targets of the relevant country. The most widely accepted regulatory objectives are:

- Promote access to basic telecommunications services at least by shared use of devices (i.e. public pay phones, telehouses etc.), by the implementation of universal service, that is to assure for all households access on affordable prices
- Foster competitive markets, which promote the efficient supply, the widening choice, the quality improvement, the price decrease of telecommunications services,
- Prevent abuses of market power, where competitive markets do not exist or fail,
- Create a favourable climate to promote investment to expand telecommunications networks,
- Raise public confidence in telecommunications markets through transparent regulatory and licensing processes,
- Protect consumer rights,
- Promote increased telecommunications connectivity for all users through efficient interconnection arrangements,
- Optimise use of scarce resources (frequency, numbers, right of way),

Telecommunications regulation depends on the legal system, on the economic development, on the culture of the particular country. Considering the common features of the individual regulatory solutions models of regulation can be constructed which are expedient to investigate the most important principles of the evolution of the telecommunications regulation. In this way three historical models of telecommunications regulation can be recognised:

- the period of natural monopolies,

- shaping of a competitive environment,
- preparation for the Information Society.

8.1.1. Period of natural monopolies

In the United States the monopoly position of the telecommunications (in those days mainly telephone) service providers was dominant till the end of the sixties and in the EU countries it was till the end of the eighties. Depending on the owner structure of the service providers two types of regulation models can be traced: the American and the European. In the American model the telecommunications service provider (AT&T) functioned as a monopoly, aided and in the mean time regulated by the state. In the European model the telecommunications service providers, generally under common management with the posts, functioned as state owned companies (the PTTs), and the state, as owner, tried to achieve its regulatory objectives.

For a long time the monopoly regulatory model was deemed to be satisfactory both from economic and from social aspects. The economic reason of maintaining this model was the belief that exploiting the benefits of economies of scale would maximise the efficiency of telecommunications, a natural monopoly by avoiding the duplication (eventually multiplication) of facilities.

The monopoly operators, working according this model, fulfilled the governments' social requirements of universal access, by undertaking *supply obligation.* The difference should be stressed: it didn't mean *universal service obligation,* which is defined as access to anybody to specified services on affordable prices. To fulfil the supply obligation, instead of using the affordable prices of the universal service obligation, the more advantageous prices for the residential users were made possible by cross-financing (pricing international and long distance tariffs high above costs and thus pricing subscriptions fees and local tariffs below costs).

It was esteemed an advantage of the American model, that the private monopoly attracted all the capital needed for the development of telecommunications, freeing the government of doing this from the taxpayers' money. The state regulation of public utility services (gas, electricity, water etc.), implementing the political objectives of the state by regulation, had been a long tradition in the United States. The state regulation of private operators began already

at the end of the 19th century, primarily to assure the country's security and defence potential.

More thorough economic investigations, however, revealed numerous failures of the monopoly based telecommunications regulation model. One of the principal failures was the „regulatory capture", which means, that neither the regulator could conclude optimal decisions against the state-aided monopoly, nor the state owned operator could make economically optimal decisions against the continuos political interference. Therefore the monopoly operators' economic efficiency and ability for innovation deteriorate. As a consequence, the monopoly operator is not interested in the introduction of new devices and services, although the digitisation and the evolution of semiconductor technology offer a wide choice.

8.1.2. Shaping a competitive environment

The American model

In the United States the monopoly model faced serious challenges from many sides. The vertical monopoly - which unsurprisingly evolved from the service provider AT&T, and the equipment manufacturer Western Electric – hindered the introduction of data processing, the new Value Added Service provision possibility. This offended the interests of big users (i.e. banking, transport etc.) who realised that this monopoly structure deprives them of providing new services and hampers their cost savings. AT&T, the service provider in monopoly situation also impeded the introduction of remarkable innovations affecting telecommunications equipment (new technologies of wireless transmission, of switching and of customer premises equipment, satellite technologies). The American telecommunications policy didn't aimed at basic transformation of the sector, but intended to reach a gradual elimination of the telecommunications monopoly. The American regulator, the FCC (Federal Communications Commission), between the end of the sixties and mid seventies, under the pressure of the market players to cease the monopoly restrictions, allowed for more and more participants to deliver telecommunications equipment, for alternatives to provide special long distance, satellite and data communications services. Thus the American telecommunications policy gradually shifted from unconditionally defending monopoly, towards reviewing the justification of the

defence of the monopoly. The result was in 1984 the divestiture of the AT&T by a judicial decision based on the American antitrust law; in the course of the trial the AT&T could not prove that the monopoly restrictions serve the interest either of the sector or of the consumers. (It is not astounding that among the protagonists of the divestiture Motorola and Hughes were to be found, who were deeply interested in the barrier free marketing of new telecommunications equipment, and satellite communications equipment respectively.)

Based on present experiences it is an obvious recognition that in the United States the erosion of the old, monopoly service structure was caused by a process referred nowadays as convergence.

In the United States the Telecommunications Act of 1996, which entirely liberalised the market, with coming into force of the relevant FCC decisions, had an effect on the players of the market. Its most important measures: mandated the spread of interconnection in a wide range, mandated the unbundled resale of telecommunications services. The Act allowed the so far banned provision of long distance services for the incumbent local telephone operators, the regional Bell operating companies (RBOCs) established by law together with the divestiture of AT&T if they fulfil a rather long check-list of requirements. It includes among other things: obligation for interconnection, making possible the resale of all their services, upon request for reimbursement a non-discriminatory, unbundled access to their local loop, transmission and switching facilities and rights-of-way, further access to their database and signalling system.

European model

Countries functioning following the European model (with the exception of the Northern European countries) faced the problem of the PTTs' poor efficiency. It was typical, that the PTTs were active payer into the state budget, or the deficit of the postal services was cross-financed by the telecommunications income. They provided services for the public, but it was not a Universal Service. It was common that the management of the PTTs' was not elected on the ground of professional competence, but it served as a well paying post guaranteed for executives appointed to political considerations. In this region it was also inevitable to transform the bureaucratic, by the politics heavily influenced PTT administration into an efficient,

market driven organisation based on new telecommunications policy, on new regulatory principles and solutions.

In 1984 the European Union initiated to create competitive environment for telecommunications aiming at the driving the sector on a new, predetermined development trend. Some of the basic elements are: elaboration of common standards; carrying out joint, Community wide research involving operators and industry; launching programs to develop the telecommunications infrastructure of the most underdeveloped regions financed by the structural funds, working out harmonized European position for the international telecommunications fora.

The next step of the European Union was the issue of the Green Book of 1987 by the Commission, initiating a Union wide discussion on the telecommunications regulatory environment aiming at the shaping a telecommunications regulation which fulfils the requirements of a harmonised European internal market. As a result of the discussion the Council Resolution of 1988 indicated the cornerstones of telecommunications policy European Union:

- creating an open, Community-wide market for terminal equipment by the mutual recognition of the type approvals;
- creating progressively an open, common market for telecommunications services to promote the evolution of Europe wide services according to market requirements and appropriate social needs. As one of the essential elements the Community-wide network integrity should be ensured working on the principle of full interconnectivity between all public networks concerned in the Community;
- narrowing the gap between the more prosperous and the less-favoured areas by providing these areas with high technology telecommunication networks;
- continuing Community measures regarding common standards in the telecommunications sector, this approach should include definition of common tariff principles;
- developing a common market on which telecommunications administrations(!) and other suppliers can compete on an equal footing; to this end, the following are particularly appropriate:
  - clear separation of regulatory and operational duties,
  - application of the relevant - notably competition - rules, to telecommunications administrations and private providers,
  - the creation of a transparent fiscal environment,
  - full opening of the markets for telecommunications supplies;

- stimulating European co-operation at all levels, particularly in the field of research and development, in order to secure a strong European presence on the telecommunications markets,
- creating a social environment for the future development of telecommunications;

  - secure the dialogue between partners for developing social consensus concerning the transformation of working conditions and lifestyles resulting from the new telecommunications technologies,

  - given the changing requirements, steps must be taken to see that workers have the right skills, to protect personal data;
- working out a common position on satellite communications, so that this new information medium can develop in a favourable environment;
- There should be prior negotiating positions of concern to the Community in the international organizations dealing with telecommunications, in particular the meetings of the ITU and on those aspects of the Uruguay Round which cover telecommunications.

The opening up for competition the European telecommunications market was required from the Member States by two Directives in 1990: the “Open Network Provision (ONP) Directive” and the “Service Directive”. These Directives prescribed for the Member States the withdrawal of exclusive rights granted to telecommunications organisations, for service providers with special or exclusive rights to make their networks accessible on commercial terms. The Open Network Provision was accepted as the basis of regulation in the Countries of the Union. The aim was, to grant telecommunications infrastructure, which could be jointly used, on the same conditions to anybody, who would provide services using it, thus creating competition without all the new market entrants should construct a costly infrastructure again. It was also mandated that the tariffs should be transparent, and for services provided on exclusive rights cost based. The complete liberalisation of Value Added Services till the end of 1990, and of the data services till the end of 1992 was also stipulated. Voice telephony service was not opened up for competition, because the price structure of the telecommunications companies was far from reflecting costs and it was likely that the competing service providers concentrate on the very lucrative international services gaining a considerable market share only because of the distorted price structure. In the mean time tariff rebalancing was urged.

The „White Paper on Growth, Competitiveness, Employment" issued by the European Commission in 1993 placed telecommunications in the centre of the Union's policy. Subsequently, in 1994 the policy makers of the European Union, to solve the accumulated social and economic troubles – following the example of the United States and Japan – accepted a new strategy with the fundamental element of building up the Information Society. From this the improvement of European competitiveness, the speed up of economic growth, the increase of economic and social cohesion, the decrease of costs and in the mean time the increase of the level of public-utility services together with it the improvement of the citizens' quality of life was waited. The "Bangemann Report", supported by the 1994 Korfu meeting of the Council of Europe outlined the way to the Information Society. It stated that Europe should break with the policy based upon fundamentals valid before commence of the information revolution. The key to the rise of new markets is a new regulatory regime allowing full competition. The new regulatory regime should be calculable in order to make strategic planning and investment feasible, further it should support and defend competitive environment. Prominent condition is the freeing of operators from political restrictions.

An important milestone of the efforts, which intend the establishment of the European Union's information policy, the full liberalisation of telecommunications and the assurance of the crucial regulatory environment, was the Resolution of the Council of Europe in 1993,. According to which voice telephony service should be liberalised in the majority of the member states till January 1, 1998 (transitional period was allowed for Spain, Ireland, Greece, Portugal and Luxembourg).

The Council Resolution adopted in 1994, which can be viewed as a guide towards liberalisation, states that the necessary regulatory framework in order to ensure effective liberalisation, shall set up common principles ensuring, inter alia:

- the provision and the financing of universal service,
- the establishment of interconnection rules,
- the setting up of licensing procedures and conditions,
- comparable and effective market access,
- fair competition.

It should be noted that solutions similar to the American and European models are to be found in the General Agreement on Trade in Services (GATS). This WTO

document on telecommunications signed in 1997 by 69 countries, carrying 90% of the World's total telecommunications traffic, committing themselves to liberalise the telecommunications market was the first multilateral commercial agreement declaring the following binding interconnection rules:

The interconnection with "Major Supplier" (it can generally be assumed to refer to operators with a dominant position vis-à-vis essential infrastructure or market share) must be assured

- at any technically feasible point in the network,
- in a timely fashion,
- on non-discriminatory and transparent terms (including quality and rates),
- sufficiently unbundled to avoid charges for unnecessary components,
- at non-traditional interconnection points if requestor pays charges,
- in a procedure made public for the major suppliers.

The transition from the period of natural monopoly to the creation of a competitive environment takes a longer time and needs regular corrections. In the American model the corrections could be realised in the process of bringing the decisions of the FCC to the court, due to the verdict of the judiciary FCC modifies its decision.

In the European model the practice is that the Commission in all of the member states regularly reviews the implementation and enforcement of the regulatory measures aiming at liberalisation and reports on the findings. Up to now six such reviews took place and has been made public as: "Report on the Implementation of the Telecommunications Regulatory Package" in May and October 1997; February and November 1998; July 1999; May 2000 and the seventh at the end of 2001. Generally the consequence of these surveys was the revision of the existing regulatory means (directives, decisions, recommendations, resolutions), or new legal means were issued.

The most remarkable changes in the transition period to full liberalisation could be found in the Directive of 1997, which for an asymmetric regulation introduces the concept of operator with Significant Market Power (SMP). (An organisation is presumed to have significant market power when it has a share of more than 25% of a particular telecommunications market, but NRAs should take into consideration other features when determining market power). SMP operators shall publish, by the

NRA accepted, reference interconnection offer (RIO), in relation to interconnection accounting separation is prescribed. The directive gives a mechanism for sharing the unfair burden on an organization under universal service obligation.

It is informative to overview the key regulatory elements which are emphasised in the review of the Commission. In the Sixth Report these are the following: changes since the last review; National Regulatory Authority and appeals; licensing; interconnection (RIO); local access (local loop unbundling, wireless local loops,); universal service/consumers/users; mobile services, including third-generation and roaming (3G: UMTS, TETRA; mobile virtual network operators); tariffs; cost accounting; leased lines; numbering (carrier selection, number portability); rights of way; data protection; internet;

8.1.3. Regulation in the Period of Preparing for the information Society

The guidelines of a telecommunications policy pointing beyond the liberalisation of 1998 and intended to initiate a widespread debate, were laid down in the "Green Paper on the Convergence of the Telecommunications, Media and Information Technology Sectors, and the Implications for Regulation" issued at the end of 1997.. One of the essential statements of the publication was that the development of technology resulted in the possibility of providing the same services using different transmission methods. Traditionally the regulation depends on the transmission methods (i.e. voice telephony, broadcasting, Internet) this results in the contradiction of regulating the same service in a completely different manner. The paper, which was open for discussion proposed to put an end to this anomaly. It was anticipated to reach this goal by a simplified system of directives, consisting of a Framework Directive and four specific directives (on authorisation, on access and interconnection, on universal service and consumers' right, and on data protection in the telecommunications). Simultaneously it rendered the possibility to update the system of telecommunications directive, which turned to be rather chaotic because of the lot of modifications and amendments. The most important principles of the new directives, which should be implemented from May 2003, are:

- liberalisation of the "last mile" of the telecommunications market by the local loop unbundling, to make possible a cheaper and faster Internet access,

- introducing a flexible juridical mechanism which can follow the development of technology and of the market, pushing back the regulation when a competitive market evolved,
- creating equal opportunities in the entire territory of the Union by facilitate market entry and by the harmonisation and strong co-ordination of the regulation,
- adjusting the regulation to an intensifying competition, restraining the effect of on market power based regulation to operators in dominant position defined by the competition law of the Union
- maintaining universal service obligation to avoid exclusion from the Information Society

The preparation and political backing of the new system of directives was based upon the „*e*Europe initiative", released by Mr. Romano Prodi in December 1999 and approved by the Europe Council meeting in Lisbon, in March 2000.. This document, prepared in the time of Internet euphoria ascribed Europe's backwardness (vis-a-vis the USA) to the costly telecommunications (lack of competition), to the weakness of "digital on-line literacy", to the insufficient of enterprising spirit.

As a solution the advantages for the youth offered by the digital age were proposed to reach by: cheaper and faster Internet access; accelerating the introduction of e-commerce; faster Internet for researchers and students; secure networks and secure access using smart cards; venture capital for the small and medium size enterprises; making the new electronic technologies to disabled people; on-line health care; intelligent transport systems; government on-line: electronic access to public services

References

[8.1.1] Melody W. H. :Policy Objectives and Models of Regulation. in Telecom Reform Principles, Policies and Regulatory Practices ed. W. H. Melody. Technical University Denmark 1997 Lyngby.

[8.1.2] European Commission: Towards a dynamic European economy: Green Paper on the development of the common market for telecommunications services and equipment, COM(87) 290 final, 30.07.1987

[8.1.3] Council Resolution on the Development of the Common Market for Telecommunications Services and Equipment up to 1992 (88/C 257/01) 1988 June

[8.1.4] Council Directive of 28 June 1990 on the establishment of the internal market for telecommunications services through the implementation of open network provision (90/387/EEC; OJ L192/1, 24.07.90)

[8.1.5] Commissin Directive of 28 June 1990 on competition in the markets for telecommunications services (90/388/EEC; OJ L192/10, 24.07.90)

[8.1.6] White Paper on Growth, Competitiveness, Employment - The challenges and ways forward into the 21st century, COM(93) 700, 05.12.1993

[8.1.7] Europe and the Global Informatiom Society: Recommendations to the European Council. High-level Group on the Information Society, (Brussels May 26. 1994)

[8.1.8] Council Resolution of 22 July 1993 on the Review of the Situation in the Telecommunication Sector and the Need for Further Development in that Market, 93/C213/01; OJ C213/1 (6 August 1993)

[8.1.9] Council Resolution of 22 December 1994 on the principles and timetable for the liberalization of telecommunications infrastructures (94/C 379/03)

[8.1.10] Green Paper on the Convergence of the Telecommunications, Media and Information Technology Sectors, and the Implications for Regulation. COM(97) 623, 3.12.97.

[8.1.11] Commission Directive 97/33/EC of the European Parliament and of the Council of 30 June 1997 on interconnection in Telecommunications with regard to ensuring universal service and interoperability through application of the principles of Open Network Provision (ONP)

[8.1.12] Commission Directive of the European Parliament and of the Council on a common regulatory framework for electronic communications networks and services COM (2001) 380. Brussels 4. 7. 2001

[8.1.13] Commission Directive of the European Parliament and of the Council on the authorisation of electronic communications networks and services COM (2001) 372

[8.1.14] Commission Directive of the European Parliament and of the Council on access to, and interconnection of, electronic communications networks and associated facilities. COM (2001) 369

[8.1.15] Commission Directive of the European Parliament and of the Council on universal service and users' rights relating to electronic communications networks and services. COM(2000)392; COM(2001) 503 Amendment

[8.1.16] Commission Directive of the European Parliament and of the Council on the processing of personal data and the protection of privacy in the electronic communications sector. COM(2000)385

[8.1.17] Regulation of the European Parliament and of the Council of 5th December 2000 on unbundled access to the local loop. (2000/ 185COD)

[8.1.18] Proposal for a Decision of the European Parliament and of the Council on a regulatory framework for radio spectrum policy in the European Community, COM(2000) 407

[8.1.19] eEurope, An Information Society For All Communication on a Commission Initiative for the Special European Council of Lisbon, 23 and 24 March 2000

[8.1.20] Commission Directive 2000/31/EC of the European Parliament and of the Council on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market.

8.2. Creating competition in the telecommunications markets

Zoltán Pápai, author

Ferenc Bánhidi, reviewer

8.2.1. Basic economic characteristics of the telecom industry

In many respect the telecom industry does not differ from other industries but in some very important ones is really different. Being a network industry, it requires a fundamental element, a network, which physically links every parts of the vertical and horizontal structure. Moreover the telecom network is a two-way network that is it provides any-to-any connectivity, a communication link from A to any B, and vice versa. The physical network and two-wayness critically determines the economics of the industry.

In order to reach the customers, any player needs a network or access to networks, through which is capable of reaching the customers. The network is a huge physical asset, and those who want to have it, have to invest a lot of money in advance before providing the service. This investment is a sunk cost because almost nothing from the physical assets can be used in other industries, for producing other products. For the high sunk cost, entry is costly, and the investment risk is higher than in the case of other industries, where the assets are convertible and can be easily diverted to other uses. For the investors the profit opportunities give the sign and grant the motive for investment. If the prospective compensating gains compared to other industries are low, the investors walk away. Without earning opportunities there will be no investment. If there is no investment, there will not be technical and service and quality improvement, perhaps deterioration will happen. High fixed costs create entry barriers, moreover the duplication of the network facilities in many times prove to be undesirable.

Accordingly, in case of natural monopoly elements of the infrastructure, it is economically more rational to seek an alternative to the duplication. Granting access

to those bottleneck facilities helps the entry into the prospectively competitive parts or activities in the service chain. Unbundling and access rights go hand-in-hand. There is a widely shared belief among industry experts, that careful design and support of them can assure that as a blissful result of the emerging competition, the public gets wider choice, lower prices and better quality. In case of supporting entry by access to unbundled facilities, nobody can leave out of consideration without impunity that there is a need for maintaining the incentives for investment and innovation, otherwise we lost in one side what we gain in the other.

The interconnection of the networks is not a requirement *per se*, it is rather an economically rational answer to exploit the gains from network externalities. Network externalities exist because there is a gain for everyone if a new member joins to the network. Moreover the greater the network is, the higher its value to each of its members. Consequently, it is socially desirable that the separate networks should be interconnected in order to promote the social welfare. Some regulation may be helpful in order to promote these interconnections, because there are incentives for some players to reject the offer of others for interconnection. The reason is that the different connected networks are competing with each other in many respects, and one of these is the difference in size. The larger the network, the more attractive it is to the new subscribers, who have not chosen provider yet. Accordingly, the size is a competitive advantage, which can be used, or in some cases misused, in order to gain more customers, and increase further the initial advantage.

8.2.2. Meaning of competition

Promoting the social welfare, everybody has to keep off the illusion of omnipotence. The economic activity is so complex, that it is hardly imaginable that there is a real chance of reaching better result by direct control and social engineering, than by the market mechanism. According to Hayek, a famous economist of the Austrian school, the market activity is a discovery process, and hardly substituted and operated better by any planning mechanism or agency. For this, it seems a good idea to build on the forces of competition even in those industries, which formerly were ruled by monopolies. The implementation of this demonopolization process is usually called liberalization. The goal of this transition is to initiate competition in order to improve the welfare of the society. We must make

some caveats before the detailed investigation of the potential uses and opportunities of the competition.

Competition is not an end in itself, it is solely the means, which is at very many times but not overall the best means to serve the public interest. Competition can be harmful in some cases, when it is economically not desirable, and its results are produces less gain than harms or losses[1]. Workable competition cannot be created with force, or without incentives to the expected players. The regulator has to stay behind the scene, and have to abstain from too much direct involvement into the market process. The workable competition is even better than any regulator. Therefore we must leave, promote and facilitate competition by cutting back and relaxing the barriers to entry, by mitigating the effects of market power, by granting fair access to bottleneck facilities, etc.

Creating winners and losers is a normal accompanying feature of competition, and that is the price paid for service and economic efficiency improvements. Competition is impersonal and blind, and gives the glory to the best and most effective, and does not regret those who have lagged behind. Losers are the cost to the system, are the sacrifice from the point of the society in order to reach the gains.

The competitive outcome is the optimum in welfare respect if the costs are non-decreasing with the output. The dynamic approach puts the emphasis on the process and stresses the dynamic effects, namely the innovation, quality improvements, and technological development. The static model is easy to understand, and relatively easy to work with, nevertheless its scope is very limited, especially in markets, which are complex in nature and where the technological or the demand changes are so fast. Currently, in the era of fast technical development, the telecommunication sector and especially some of its peculiar segments are rather characterized by fast changes and a lot of innovation than the traditional utility business, consequently the dynamic approach is more relevant, or at least it is better using it in parallel with the narrowly focused static concept.

[1] Competition is typically harmful when an activity is characterized by the natural monopoly.

8.2.3. Preconditions of the competition

Competition is possible end desirable in those segments of the telecom industry where the activity is not a natural monopoly, where there is a living place for other players without net loss to the society (at least in the long run). The telecom industry is a set of complex, linked, vertically and horizontally related markets. Some of these markets are really, or at least potentially, competitive like the data communication, international and long distance voice telephony, mobile telephony. There are segments in the communication infrastructure, where the duplication of facilities is not economical, so we expect that there would be no more than one player providing the service. Currently the local loop is a facility for which there is no reason to duplicate it. The same is true for the cable TV infrastructure. Nevertheless even in the case of these services there may exist opportunities to bypass these bottlenecks by using alternative technologies. Even if the telephone infrastructure was optimized for providing the voice telephony services it is now possible to distribute video signals on the telephone wires, and it is true in the reverse, since the video distributing network can be capable of providing telephony services. Currently the possibility of technical substitution is not necessarily involves that this substitution will happens on the demand side. If the price (cost) and quality of a service makes this substitution opportunity undesirable for it is not economical for the customers to consume, it is not a real substitute on the market.

During the monopoly era every service were told or thought to be monopolistic. Thanks to the technological development dissolution of the artificial fog about the monopolistic nature of all the horizontally and vertically related services, nowadays it is easier to differentiate between the services with respect to the degree of potential competitiveness. Those services, which have a low degree of potential competitiveness, are proper target for regulation in order to protect the competition in the connected market, from the players who have market power or monopoly position in the non-competitive segments.

The local loop is now such a part of the network, which is very important for reaching the costumers, and hardly can be bypassed economically in many cases. This is a typical example of the so-called essential facilities. In case of essential facilities, one part of the product line cannot be duplicable, that is cannot be built, or

bought from but one player. Since this part is essential for providing other services, it is critical to access to it for the providers of those other services, because without access they would not be capable of doing their business at all. Therefore it is essential to the viability of competition, to ensure that the terms of access should be non discriminative for all of the players, and it is obliged for the owner of the facilities to make the same offer to any competitor like to its unit or subsidiary. It is an important task of the regulator that when guarding the public interest by prescribing fair access rules to the essential facilities, do not spoil the investment incentives for its owners. The role of the regulators, who are in charge with the task of guarding and policing the public interest, is to preserve the viability of competition in the long run. He or she has to be very careful in order to keep off mixing this role with supporting some competitors against others, by enforcing the bottleneck owner to subsidize the providers of the services to which its services proved to be really essential. The task of the regulator is not to create competition at all cost, but only to promote the workable competition on the market.

The problem of access to essential facilities is usually emerging in the case of the access to the local network which is - in many situations being the only economically feasible channel to the customers - necessary for providing other telecommunication services – like telephone calls, or data communication. A very important part of the competition supporting regulatory policy during the era of the development of competition in some markets (like voice telephony and data communication) is to handle carefully the issue of access to the essential facilities, by balancing between the need for more vigorous competition, and at the same time preserving the incentives for investment and innovation, and the economic gains from the possible scope economies[2].

Now, the local loop is a part of the network, which is very important for reaching the individual customers. Therefore it is a very important regulatory issue to make it available for use not exclusively by the incumbent telecom operator, but for the new players too. It may also be true for other bottleneck elements of the network. Unbundling is the name for such an obligation, which requires the incumbent to ensure access to the nonduplicable parts elements or functions of its networks.

[2] If economies of scope exist, then it is true that the cost of providing two or more services together is always less than providing them separately.

Establishing such a requirement needs a careful economic analysis of the situation, because if a concrete unbundling decree makes it obligatory for an incumbent to offer those parts of its networks, which could or should be duplicable, for the new players, then the regulator encourages inefficient entry. Granting the profit opportunity to the new players by hurting the incumbent, causes net loss to the society, and makes the competition artificial and harmful.

The extent of competition in any sector can be measured by different means. First and foremost the price cost margin is the thing that informs us about the strength of competition. In markets characterized with moderate degree of technological development and innovation, there are many players, and the competition is effective, this margin tends toward zero. In this case the measure of the relative price-cost margin, namely the Lerner index (P-MC/P) is a good indicator of the degree of competition. The higher the index, the higher are the distortions, and the weaker the competition can be. However it is difficult to measure the marginal cost. Therefore we need other types of indicators, available for practical use and good enough for serving as a basis of regulatory analysis and decisions. These indicators are mostly related to the market structure, and not solely, but together can give enough information about the situation of the market. These indicators are:

- Number of firms on the market
- Individual market shares of the players
- Number of buyers compared to the number of players
- Concentration measures like C4, C8 or HHI

C4 and C8 are the cumulate % of the market share of the first 4 and 8 firms respectively. The HHI is the Hirschmann-Herfindahl index, which is the sum of the square of the individual market shares of the firms operating in an industry.

This indicators and other information helps to judge the market situation with relatively low probability of error. Our judgment of course is based on the assumption that there is a relationship between the market structure and the behavior of the firms on the market, i.e. the market performance. The performance of the market depends on the behavior of the firms. This is dependence is reciprocal, since the behavior of the firm is influenced by the prevailing market conditions.

The entry and exit conditions, which might exert considerable influence on the costs of the players, are very important determinants of the market structure. In the

telecom sector we usually meet with high entry cost, and therefore effective barriers to entry in those segments where the duplication is impossible, uneconomical, or undesirable. We should know of course that it is not the case for very many segments, like international calls, and data communications where the infrastructure investment is relatively low or easily transferable, and the prices are high enough for inducing the potential entrants to take the risk. If the assets cannot be transferable when someone wants to leave the market, it works as a typical barrier to exit. In this case, large part of the investment and thereby the costs can be sunk. Segments with high sunk cost are not too attractive for risk averse players. Only high expected rate of profitability could induce the investments, and attract the investors into these segments.

In many cases not the cost conditions are what prevents that new players come to the market. There maybe legal barriers to entry, like special licensing procedures and requirements, corporate laws, special permit requirements, or at one extreme of the scale: exclusivity, i.e. a legally granted and protected monopoly.

8.2.4. When the competition may be harmful

Beyond the case of natural monopoly there may be situations when competition in a market is not good for the society, even it can be definitely harmful. If the price-cost margin is high enough to attract many new entrants, even if their cost is higher than that of the incumbent. This result maybe harmful in two respects:

- The high priced services are produced at higher cost by the entrant than the incumbent, so the resources of the society, are used inefficiently.
- The profits from services with artificially large margin are used to finance the loss, resultant from those services, which are priced below their average cost (as it is usually the case of under priced access and local call services). The fast exhaustion of the cross-subsidizing sources, which are necessary to maintain the system sustainable, may cause a hasty collapse of the incumbent operators, who are obliged to bear the burdens, and financial consequences of the politically induced biased price system.

The “cream skimming”, or “cherry picking” is absolutely rational from the part of the newcomers. So it is the task of the regulator to keep the balance, by preventing or minimizing the chance of occurrence, or at least the harmful effects of this situation. A well-designed tariff rebalancing policy is the best way to obviate the occurrence of the inefficient entry.

It is against the common sense, but sometimes happens that the prices set by one of the players are too low, that is low enough to discredit, and eventually terminate the competition in the long run. If the low price is the sign of the cost subadditivity which is characteristic of the natural monopoly, there is nothing to be and should be done with this phenomenon. However, in some cases it is just the sign of an aggressive pricing of the dominant player in order to supplant its challengers. In this case the company is sets and keeps the price of one of its product considerably below the average cost of this service. Even if it causes loss for the company, it might have good strategic reason to follow this policy. It may have some source, for example some other service with monopoly market, generating monopoly profit, from which it is capable of subsidizing this pricing policy. As a result of this so-called predatory pricing, it can gain greater and greater market share, eventually eliminating its weaker rivals from the market. At the end, when the rivals will have crowded out of the market and the predator remains alone, it will lift the price of the service well above its cost, and enjoy the fruits of the obtained monopoly. This predatory pricing behavior, with the consequence of destroying the competition, is really harmful for the society. Therefore the regulators have to be vigilant guardians of the customers' long-term interest, protecting them from such predatory behavior. The real problem with this predatory behavior is, that it is very difficult to detect and prove it, and even after the detection of these consequences of the behavior, to prove the predatory intention of the firm.

8.2.5. What kind of competition we prefer?

There is a question of what kind of competition we would like to see blossoming on the telecom markets, in order to promote the public interest. It is a real question from the very beginnings of the telecom liberalization. One extreme opportunity is to initiate, support and enhance the competition between network infrastructures; the other is to create competition between service providers, using the same infrastructure. Both ways have their pros and cons, and neither seems unequivocally superior to the other. We list the pros and cons in table 8.2.1.

	Pros	Cons
Infrastructure competition	Alternative/parallel infrastructures are good for competition in the long run It may enhance the technological development, and the competition between alternative technologies, thereby granting further benefits to the society in the long run Requires commitment and provide incentives for market achievements, since it requires investment in non easily transferable assets, thereby increasing the exit costs	It takes a lot of time to build the infrastructure, and some more time to realize the gains for the society It needs some protection for the new entrants, during a not well defined incubation period. But this protection may destroy the incentives for innovation and efficiency Some parts of the network might be natural monopoly The new players might not enjoy the same economies of scale, scope and density like the incumbent It assumes well designed policy with regard to interconnection obligations
Service competition	Entry and exit are less costly than in the case of infrastructure competition, and the sunk cost is relatively low May attract more players than the infrastructure competition Relatively easy to establish spectacular competition in some market segments, and show some good news relatively soon for the society Promotes innovations related to service and delivery (marketing, billing, customer care, etc.)	It requires well designed, nondiscriminatory access rules and terms It requires some solution to give incentives to the infrastructure owners for maintaining their interest in the provision of good service quality, and research and technological development There is a bias from the part of the newcomers toward cream skimming, making fortune quickly and shaking the dust from their feet It requires protection against the incumbent's propensity to price squeeze[3] It requires organizational, or at least accounting separation and regulatory oversight of the vertically linked business activities at the incumbent

Table 8.2.1

8.2.6. Key regulatory issues in the era of competition building

In order to protect and enhance the competition, the regulator has to take care of the assurance of nondiscrimination in access and interconnection rules and terms, and has to bloc any anticompetitive practices stemming from any involved parties. In addition, it has to keep off any needless interference into the market process, and refrain from determining the market outcome in place of the market itself. The ability of self-control is one of the highest valued merits of a regulator.

In the era of competition building the role of the regulator and the content of regulation is changing in comparison with the natural monopoly era. When the competition in the nascent phase, the need for industry specific regulation is increasing, and then giving its place to the classic competition regulation, when the

[3] The definition of the price squeeze can be found in Chapter 8.3.

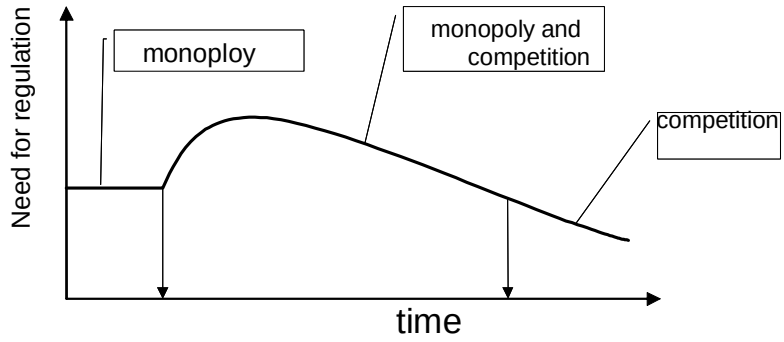


Figure 8.2.1 The change of need for regulation [Europe's Network Industries: Conflicting Priorities, CEPR]

competition is strong enough. Figure 8.2.1 shows the changes in the demand for sector specific regulation after liberalization.

In the liberalized markets the most important new regulatory tasks are the followings:

- Inhibiting the misuse of market power
- Assuring the nondiscriminatory access to essential facilities
- Elaborating guiding principles, and methods for access and interconnection negotiations
- Taking part in dispute resolutions
- Checking and barring the predatory pricing activities
- Carefully following and analyzing the market process, and interfering only in case of there is something against the public interest

After opening up the telecommunications markets the content of the regulation changes considerably. The regulator has to search for a new optimum between the decisions influenced by prescribed regulations and truly independent decisions of the firms on the market. Figure 8.2.2 shows the hypothetical relationship between the decisions imposed by the regulator, as a portion of the total, and the welfare effects of this setting.

It is shown, that in the optimum under competition the regulator do not have to be involved too much in the decisions of the market players, though at the same time this regime results in higher welfare level. It is also seen, that if the sector regulator did do the same way as in the era of monopoly regulation, the welfare would decrease even lower than in the former optimum. The regulatory challenge is that how to jump up to the optimum after liberalization, without falling down to the valley

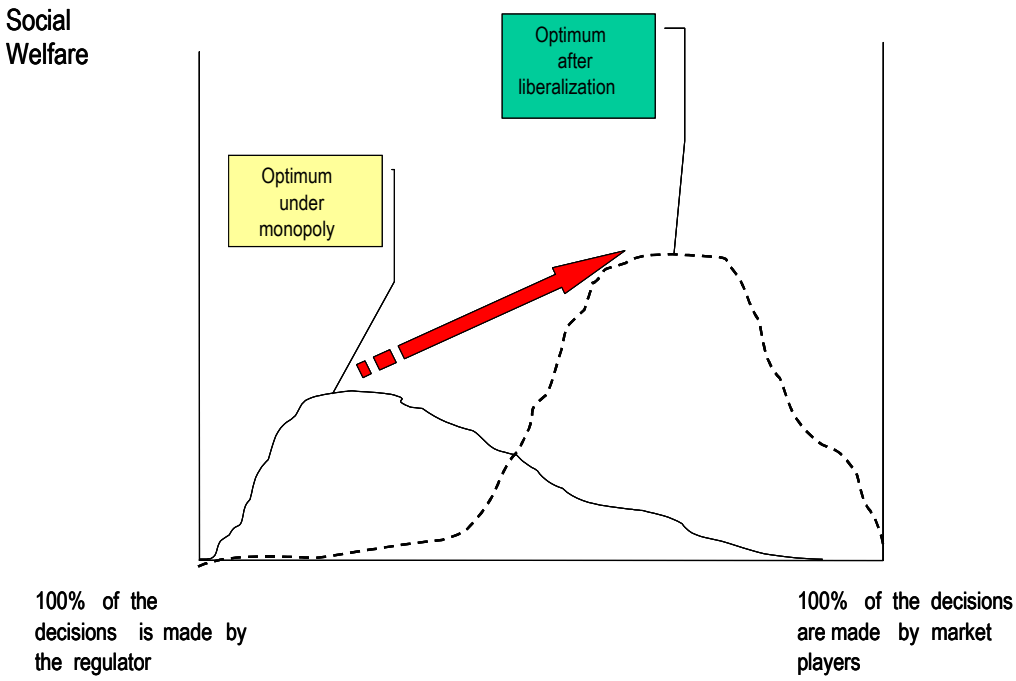


Figure 8.2.2 Distribution of decisions made by market players and regulators

between the optimum peaks. If it falls down, its fate would be the same as if it did choose the way of incremental adaptation, the so-called “muddling through” strategy. It can reach the peak after tiresome uphill passage.

In order to fulfill these challenging regulatory tasks, the regulatory institutions, and the regulators themselves have to be changed. They have to be open-hearted supporters of the market process without any bias toward any of the players. The characteristics of the new regulator and regulatory process should be the following:

- Independence
- Openness
- Transparency
- Impartiality
- Accountability
- Professional competence
- Non-interventionism

The new telecommunications regulatory institutional setting should become rather like the competition authorities, than the traditional regulatory institutions. And the progressive regulators always have to keep in mind that their task is to leave the

competition follow its direction, and accept the outcome, instead of determine the results as a substitute for the competitive process. The worst what a regulator can do is to interfere constantly into the market process. The good regulator should overcome its strong propensity to micromanage.

References

[8.2.1] Berg, S. V. – Tschirhart, J.: Natural Monopoly Regulation. Cambridge University Press, 1988

[8.2.2] Bergman, L. – Doyle, C. – Gual, J. – Hultkrantz, L. – Damien, N. – Röller, L.-H. – Waverman, L.: Europe's Network Industries: Conflicting Priorities, Monitoring European Deregulation 1: Telecommunications. Centre for Economic Policy Research, 1998.

[8.2.3] Intven, H. (szerk.) : Telecommunications Regulation Handbook, InfoDev, The World Bank, 2000

[8.2.4] Kahn, A. E.: The Economics of Regulation: Principles and Institutions, MIT Press 1988

[8.2.5] Peltzman, S. – Winston, C.: Deregulation of Network Industries: What's Next? AEI-Brookings Joint Center for Regulator Studies, 2000

[8.2.6] Viscusi, K. W. – Vernon, J. M. – Herrington, J. E. Jr.: Economics of Regulation and Antitrust (Third ed.) MIT Press, 2000

8.3. Some issues of economic regulation

Zoltán Pápai, author

Ferenc Bánhidi, reviewer

Some developed tools of the monopoly regulation may be useful and in use further in time, at least for a relatively long transitory period. This must be peculiarly the case for those activities, which still remain monopolies even in the competitive era. Therefore there is a place for applying the elaborated tools of classic price regulation for by monopoly provided services, there was a tendency, that the direct interconnection regulation gives place to a regime where the interested parties negotiate the price and technical terms of interconnection, some regulatory involvement seems to be useful, or sometimes inescapable in order to cut the Gordian knot in the disputes between the market players.

Until now we have studied the tools of the so-called economic regulation. But there may be regulatory involvement with the intention to enforce or promote other, non-economic goals. One of these, and presumably the most remarkable is the regulation of the universal service provision, from social policy mainspring. It is known, that every type of regulation have economic effects, but the regulation of universal service is unique, because it effects on the whole set of economic relations of the sector for the existence of the required cross subsidies. Providing the basic telecommunications services available for the whole society is not only a sounding social goal drawn by politicians, but an almost universally accepted feature of the communication services in the developed world. The content and financing of the provision of universal service is still a critical and strongly disputed element of the telecommunication policy of those countries where it has been established.

8.3.1. Price regulation

The regulation of end user tariffs is not a new phenomenon. It was the preeminent goal and rationale of regulation as an economically useful and politically desirable activity. However it is now a legitimate question, whether the old tools are appropriate enough for handling with new type of markets. Of course it is also a

relevant question whether characteristics of the markets in question are really new enough. At the beginning we must make unmistakably clear that there may be some need for regulation if only the competitive forces cannot work well enough to produce the economically desirable outcome. But this is only the necessary condition for intervention. The rationale is sufficient if and only if there is a significant chance that we expect better outcome with the help of regulation than what we can have without it, what is the case with every type of regulation. The problem is here the there is a temptation on the part of politicians to use regulation as a redistributive tool, as a means of hidden taxation. It has never been a good idea, but in the time of competition it may be extremely harmful, for by regulating competitive services the regulator may destroy the competition itself. If the regulation eliminates the profit from the prices of potentially competitive services, it throws the baby with the bath water by destroying the competition itself. In this case there is no incentive to enter into the market. In the case of end user services there is place for regulation for only those markets, which are relatively intact from competitive forces. In most of the cases the provision of local loop and local call services are proper candidate for price regulation, especially in rural or not densely populated areas. After opening the telecommunications markets there seems to be no rationale for regulating the prices of long distance, and international call services as well as those of value added services. The countrywide availability of *call by call* selection or the *carrier pre-selection* opportunities offers choice enough to eliminate price regulation for these services. Of course there may be some temporary period during which there is a need for some loose type of regulation with the rationale of taking care of the interest of ignorant costumers who may be ignorantly choosing the incumbent's default option.

The end user price of local access and local call services (if there is no competition in the provision of them) can be regulated with the traditional price regulation tools. These tools are called "traditional" not because they are really old enough, since the price cap regulation is barely used for less than two decades. The word "traditional" refers only to the fact that this tool was developed for regulation in the natural monopoly environment. The theoretical basis of price caps is that if a firm is getting only normal return that is zero profit with the current prices, whenever its input prices rise because of the inflation, it is only necessary to allow to rise the

output prices by the inflation rate minus a preset requirement of expected growth factor for productivity. This type of regulation usually takes the form of *Inflation index-X* regulation, where the inflation index is usually the *CPI*, that is the normal statistical measure of the consumer price index and *X* is the productivity requirement. This is of course the main formula and it sets a limit for the change of aggregate prices. It is the task of the regulator to fine-tune the formula and to fit it to the peculiar situation of the different services, having linked but different markets. In order to understand how the price cap regulation really works, we should say some words about its predecessor, the rate of return regulation, after than we discuss shortly the most important details.

The rate of return regulation had been used since many years in the US before new regulatory techniques changed it. The regulatory process is a “rate case” where the utility company applies for the approval of its price increase plan to the regulator. The reasons of the price increase is that since change has happened in the level of costs, therefore there is a need for price increase in order to keep the rate of return on the company assets on the regulatory accepted level. This technique is called “cost plus” regulation, because the regulator merely decides whether to allow or not the change, the company asked for. The regulator may investigate whether the costs were necessary and right for the production of the output, but has no right and tool for judging or questioning the technical efficiency. That is why there are problems with the incentives for enhancing efficiency of the regulated firms. Moreover the rate of return regulation has other flaws too. In case of rate of return regulation there are no incentives for innovation and efficiency enhancements, and unfortunately there is a perverse incentive to choose capital-intensive technology in order to blow up the rate base and thereby increase volume of the profit.

The most important requirement with the inflation index is that it has to be exogenous, that is it must not be exposed to the influence of the regulated firm. Consumer price index or core inflation may be appropriate for this choice. Another issue is whether the index will be *ex ante* or *ex post*. An *ex ante* index are known before the year starts. It may stem from inflation data from the past, or can be based on some accepted independent forecast. *Ex post* inflation will be known by the end of the year, so the regulator does not fix anything in advance, but at the end it checks whether the firm has complied with the cap or not. Each solution has some pros and some cons, however in reality the *ex ante* CPI index is used generally.

The prescribed efficiency requirement can be calculated from past total factor productivity (TFP) data or a forecast of productivity enhancements during the regulatory period. The productivity studies require good data for analysis and hard work and knowledge to process them. Outside the US and Canada the expected productivity is determined by softer methods, for example on the basis of some estimates, or by using benchmarks. In the calculation of the expected productivity measure the regulator can calculate with the difference between the firm input inflation and the inflation index of the overall economy as well as with the difference between the total factor productivity of the economy and the regulated firm or industry[4].

The price cap rule gives the firm strong incentive to work better and as a reward it can enjoy the fruits of its efforts. The use of the price cap formula is very easy when the regulated firm is a multiproduct monopoly that is to say it is a monopoly in each and every of its product markets. In case of this, the price cap regulation works as a pure profit regulation system. Regulatory problems emerge when some of the markets are competitive while others are not. Considering the competitive products, there is no need for regulation, but there is a need for regulation for the still monopoly products. Moreover regulation is needed not only for protecting the captive consumers, but in order to protect the fellow competitors in the competitive markets. This protection is needed because until the firm can earn profit in the regulated market, it has the opportunity to trade on lower than cost (called *predatory*) prices on the competitive market, thereby crowding the competitors who is lacking this source of cross-subsidy cross out of the market. So the regulator's responsibility is even greater in the liberalized regime, than in the monopoly one. The regulatory requirement of accounting separation is a supporting rule for preventing the regulated firm from making illegal cross-subsidies. Though the accounting separation is a good tool, it is not easy to operate it in a situation when inherent special economies exist, stemming from joint production.

A very important feature of the price regulation is that the regulator may set up other constraints for different groups (called baskets) of services. One of the reasons for this is that the expected productivity may be significantly different for some regulated products than for others. Another reason may be that there is an

[4] In order to understand the concrete application, see Bernstein and Sappington (1999).

acceptable regulatory goal to rebalance the historically distorted tariffs. This rebalancing is important, for only tariffs without cross-subsidy constitute and send adequate signals to the consumers and would be competitors as well. Because the one step rebalancing policy seems to be lacking any political reality, a path of rebalancing can be prescribed with the help of separating and differently regulating the prices of predetermined baskets. These are the reasons why there are some sub caps under an overall umbrella cap for the different groups of end user services. Typical baskets in the European countries) are:

- Access and local call services (monthly rental, local call services)
- Long distance telephony services
- International call services

Though price cap regulation can be applied to regulation of any services, but nobody can believe that he or she can get done with this tool without careful analysis and design of the parameters of this relatively simple formula.

Closing our short tour around the issues in price regulation, we cannot ignore new problems emerging at the beginning of competition. Quite a new problem of regulation in the competitive era is the need for regulation of prices of wholesale services. These services are services sold not to end users, but to the competitors of the incumbent firm. This wholesale product is an input for the competitors and necessary for producing retail products by both the incumbent and its competitors. If the wholesale price charged to competitors is higher than what is charged to the downstream part of the firm, it is possible to operate a price squeeze by raising rival's cost.

8.3.2. Regulation of Cooperation of the Service Providers

Interconnection is not only the key for the cooperative interoperation of the different networks, but it is really a key for the operation of competition. Moreover this issue is, above its technical complexity (see Subchapter 8.5) , the most important economic question of the industry, determines in some cases whether some players can survive at all. Without interconnection there is no competition. The proper interconnection regime is therefore the paved road to efficient competition. There are markets, where at the level of current prices there is place for newcomers with

innovative technology or any competitive advantage over the incumbent. A good interconnection regime has to allow them to enter the market and compete for customers. Whenever the decision to interconnect is the choice of the incumbent or the significantly bigger network it may resist or use its bargaining power to impose unfavorable terms and conditions on the new entrants. Of course not every of the new entrants are necessarily more efficient than the incumbent, so the interconnection regime does not have to support everyone, just those who are as efficient as the incumbent. The interconnection regulation is tackling with the required and acceptable economic principles of interconnection, and with the regulatory constraints on the bargaining process of terms and conditions of interconnection agreements. It usually takes the form of the regulation of the behavior of incumbent during the negotiations, and after setting the contract the regulator watches the parties compliance with it.

There are three types of the regulatory approach and model to the interconnection problem around the world:

1. The whole issue of interconnection is regulated, and the regulator sets the prices and checks the compliance of the parties (Interconnection, IC model 1)
2. The regulator sets the rules and requirements, but the incumbent prepares an interconnection offer according to the requirements (reference interconnection offer -RIO), and if the regulator accepts it, it will be the standard basis for individual interconnection agreements with any party who asks for it (IC model 2)
3. The regulator may withdraw from the interference in interconnection issues partly or totally, leaving the decision on the whole question to the interested parties (IC model 3)

In IC model 1 the regulator makes the decision, but it has to follow the process closely and to intervene again and again in order to correct the mistakes it has committed with its previous decision. In such a situation the market may suffer from the risk of regulatory error.

In the IC model 2, the regulator only keeps the right to decide on principles, especially of costing principles of interconnection. It is by no means a hard task, but much easier and acceptable to undertake than fixing the proper costs. According to the rules, the incumbent has to reveal what it considers appropriate under the principles set by the regulator, which is under the scrutiny of the interested parties. At the end it may happen that the regulator has to approve the incumbent's offer,

though it has the right to refuse it, if it seems to be inappropriate, and ask for a new one. In vain in this model the regulator is still in decision making position but its role mirrors its low-key position with respect to information. This approach seems to be closer to optimality in case of the asymmetric information than model 1. The emphasis in model 2 is on the fairness and appropriateness of the process, whereas in model 1 on the appropriateness and correctness of the data. This model 2 is widely used in many developed countries, especially in the EU, and this is applied in Hungary as well. In this model the regulator sets the principles and closely oversees the process, and if it is necessary arbitrates between the parties, and in order to prevent the tie, decides.

In IC model 3, the regulator does not play any role, or at best it works as moderator. Everything is left to the parties who have to negotiate the whole issue of interconnection. This seems to be the most liberal regime with respect to the question. Nevertheless it is presumably the slowest of all. Nothing prevents the incumbent from exercising its market power during the bargain. The advantage of this situation, if there would be agreement at all, that it had been set by mutual voluntary agreement. The problem with model 3 is that the chance of occurrence of such a consensus is hardly significant. In most of the cases, the story ends at the court. At the first sight it seems to be the most market friendly solution, but it is always in the interest of the incumbent. In any case, why should we think that a judge at the court, who is not educated in the peculiarities of interconnection, will be able to decide better than the trained but uninformed regulator. Such a light-handed version of interconnection regulation usually results in the delay of entry. However if any player has reached an agreement, it follows from the non discrimination principle, if it is required to apply, any third party has the right to ask interconnection with the same terms (if it is applicable). This model has been applied in New Zealand.

After overseeing the institutional solutions, let’s see the principles applied or may be applicable in interconnection issues. In order to support the development of the competition in the telecom sector the interconnection contractual terms should be (a) nondiscriminatory and (b) transparent. The requirement of avoidance of undue discrimination is a guarding principle for new entrants and smaller players who may suffer from an incumbent’s rejection or unfair treatment when they want to interconnect with this incumbent network. However it does not follow from this

principle that everybody should get the offer with the same terms with no respect to the relevant differences of size or network infrastructure background.

The transparency requirement supports the contract preparation process by letting know a third party, who also wants to make a contract with the incumbent the relevant terms and conditions offered to others. There should be information available on the market about the terms and conditions of the contracts, to make the contractual process faster and the competition more effective.

Many times the incumbent operators have so strong position that even the transparency and nondiscrimination requirements would not be effective enough to balance this advantage. Therefore many times asymmetric rules are needed in order to keep competition alive. Typical form of a burden on the incumbent is the obligation to publish an interconnection offer, which is reviewed and should be approved by the regulator. This interconnection offer serves as the incumbent's default offer for anyone who asks for interconnection.

Many times the obligation to interconnect is a rule according to the law in force. Though if there is an obligation to prepare an interconnection offer there is no need for establishing the obligation to interconnect, because the parties should have discretion to decide whether under the given terms they want it or not. The universal interconnection obligation, which prescribes an interconnection obligation for every player, is a severe excess burden on the non-incumbent operators, and it spoils the incentives to build larger alternative networks.

The views about the optimal balance between the conflicting goals (that is preserving the investment incentives and promoting competition is reflected in the different proposals of interconnection pricing. The proposed options are the followings:

- Interconnection charges should be cost oriented. But cost orientation does not tell how this cost is to be calculated (on a historical or forward looking basis), and what level of mark-up can be accepted (if at all). It has also been debated that what method of costing is accepted: fully distributed cost (FDC) or incremental costs. In the last decade the long run incremental cost (LRIC) approach has become the regulatory standard. Telling the truth that is which has any economic meaning.
- Another proposal is the principle of competitive parity or with another name the efficient component pricing rule (ECPR), firstly proposed by Willig, and called Baumol - Willig rule. This says that a new entrant has to pay for the access to

the incumbent the retail price minus the avoidable cost that is the lost profit. The idea behind this principle is that it is not desirable that operators, less efficient than the incumbent, enter to the market. If they are more efficient their presence in the market is socially beneficial. The other property of this principle is that only the incumbent is indifferent to the new entrance. While this rule is economically very reasonable, usually has not been applied.

There are very important technical issues concerning the interconnection, with which an incumbent can influence the chance of success of new entrants. These are:

- Number of required interconnection points (minimum and maximum)
- Facility and capacity requirements and interface specifications for interconnection
- Traffic measurement requirements
- Billing options
- Quality of service issues

Though these issues seem to be very technical at the first sight, they have serious economic consequences and may determine the success or failure of those new entrants, who have asked for interconnection.

The interconnection experiences after liberalization stemming from the practices of the western countries have showed that interconnection is the key element for opening up the telecom market. After the setup of the new interconnection regime, regulators have to follow carefully the process of making interconnection agreements and interfere only whenever it seems necessary.

8.3.3. Universal service policy

Though many times socio-political policy considerations contradict with economics, they are fundamental parts of our everyday regulatory practice. It is not only to think about that politicians have their stake in regulation, but welfare, justice, and income distribution and other issues are addressed in the process of regulation. One of the most striking issues in the network industries is the universal service policy. The original concept had emerged in the beginning of the XX. century with a really different meaning. Theodore Vail, the president of AT&T told in 1907 that "*One system, one policy, universal service*". He had meant that the technical integrity of the telecommunications system is the only way to conduct the business, and it is beneficial for the whole society. Today the phrase "universal service" refers to the idea that some public utility services are very important and integrant part of the

modern life and they have become necessities, and therefore it is socially desirable that they could be consumed by anyone. The concept of course contains the requirement that the price what any customer has to pay, should be affordable. So the idea is universal access on affordable price. Without saying too much about other sectors, we should tell that this idea is not limited to the telecommunications industry, although it has developed closely connected to it.

Universal service in its nowadays form in the developed countries, typically aims to support the access of poor and rural customers to the telecommunications network and services provided through that. The arguments for the proposed universal service policies in the developed countries are usually based on the following three reasons:

One rationale behind this policy is, that everyone in the network is better off, if some more subscribers join to it. It is worthwhile to subsidize the subscription of the poor in order to reach a higher level of penetration until this extra cost is equalized by the marginal gains. It limits the validity of the argument, that the increase in the size of the network constitutes increase in utility for some one, if and only if the new subscriber is such who would call him or would be called by him.

Other argument also refers to some externality, but to a social one. This says that it is better for the society to have everyone connected and a user of telecommunication services because by this way the whole society is kept integrated, and therefore there is no division of people in and out. The idea of the information society is valuable if and only if everyone has a chance to be a part of it. Though this argument is very acceptable, it has some flaws, because it supposes without doubt that everyone wants to be a part of the idealized information society. However, the intention to promote universal service, is still remaining a positive and promising social policy concept.

The third argument refers to the human needs, and the role of the society to ensure that everybody must have his or her basic needs satisfied. We must know, that the level which is called basic, is relative, because it depends whether some more elementary needs are satisfied or not. If there are more basic needs unsatisfied (like food and shelter) it is an illusory goal and a waste of resources to try to promote the take up the electronic communications services.

Those developed counties, which are talking most about the information society and universal service policy, usually do not have the dilemma that whether it is really the most acute problem of the society in question. It might be. But those countries with respectable, but moderate telecom penetration are facing with the dilemma of whether it is really the most important and urging issue to increase the penetration with every means and without respect to the costs, or let the market decide the penetration and choose the technology of the communications infrastructure. The policy maker of course may play a very important role in supporting the evolutionary process, but should avoid interfering with it. The temptation to mimic the western countries may induce a self-deceptive policy. The real problem with misplacing of the policy goals is that it squanders the valuable resources of the society, and curtails the society to satisfy more valuable and respectable needs. Every society has its own peculiarities, and we may say, with a relatively low risk of committing an error, that there is no universal solution to universal service policy.

The core of every universal service policy is to promote the close to 100% coverage of a service, which has a penetration rate above 80%, and to give a chance to access to the basic communications services who can afford it only on marginal, but not on the average price. Under this threshold there must be market opportunities for increasing the penetration and the policy maker only has to support the process, without serious interference[5].

One of the potential failures or expected inefficiencies of universal service policies in less developed countries is that affordability cannot be separated from the general economic conditions of a country. Affordability, being a relative term, is not solely about the price, it is also about the disposable income, and consequently the level of welfare and the distribution of the income in a society.

Universal service, broadly speaking, involves averaging of rates and thereby cross subsidies between different consumer groups (business-residential, urban-rural). The difference between universal service before and after the date of

[5] The question is: what type of penetration? In case of the voice telephony services it is the case, that not each and every member of a household has his or her own fixed line. Having one or two lines (in the developed countries) is enough for satisfying the communications needs of the whole family. That is the reason why the household penetration is a more appropriate measure of the fixed telecom penetration, than the number of subscriber lines per 100 capita.

liberalization is that in a monopoly regime the internal cross-subsidy is viable and sustainable, but after the liberalization is not. In case of competition the profit making services are under competitive attacks, and as an effect of this competition their profits decline and there will be less and less internal source of subsidy. That is why external subsidies may become necessary in order to keep the universal service policy alive.

In case of competition there is a need for a default provider, who is obliged to offer the universal service to satisfy every reasonable request for the service on an affordable price in its service territory. Providing universal service for those consumers, who anyway can be uneconomical, may result in losses, though there are also advantages to be universal service provider. That is why the implementation of universal service policies may require a financial support system, in a form of an institutionalized universal service fund. According to the "*pay or play*" principle, those, who provide universal service, can be subsidized through this special fund by those players, who have no such a burden.

Universal service policy may consist of different building blocks, but usually contains the following ones:

A) Access to the telecommunication network and basic voice telephony services:

- public phone boxes
- affordable subscription offers

B) Access to other communications related basic services:

- Free emergency calls
- Directory services
- Phone Books

C) Special access opportunities for disabled persons

D) Affordable access to other services, which are considered basic by policymakers (Internet, e-mail, etc.)

The universal service compensation, if it is necessary at all, may come through a fund. Though at the first sight other market players who provide public telecommunication or public voice telephony services are those, who are obliged to contribute, and not the taxpayers, but this system is nothing else than a hidden system of taxation and redistribution. In vain, customers of some services pay for the costs of others. Moreover it is almost sure, that this system is less efficient than an

optimal taxation. The only thing why it is thought to be favorable by politicians and interest groups, that it is less transparent than the redistribution through the state budget.

The principle of the calculation of the net universal service cost is rather simple. The net cost is the avoidable cost of service[6] less the sum of financial and non-pecuniary benefits[7] of provision. It is less clear in practice than in theory, that what constitute the really avoidable costs. But one thing is sure: it is impossible to calculate the costs case by case. Regulators usually try to use cost models for net loss calculation. What is usually missing from the practice, a real cost-benefit analysis of the policy and its social effects in order to get a balanced picture of distribution of gains and losses.

There are a lot of controversies around the universal service policy in every country, and more is expected when less and less developed countries are trying to establish it. The need for keeping dynamism in the telecom sector requires careful analysis and balanced decisions in considering the technology (if it is necessary at all) and actual content of the universal service policy.

References

[8.3.1] Armstrong, M. – Cowan, S. – Vickers, J.: Regulatory reform: Economic Analysis and British Experience. MIT Press, 1994

[8.3.2] Arnbak, J. – Mitchell, B. – Neu, W.– Neumann, K-H. – Vogelsang, I.: Network Interconnection in the Domain of ONP. WIK, 1994

[8.3.3] Baumol, W. J. – Sidak, G. J.: Toward Competition in Local Telephony, MIT Press & AEI Press 1994

[8.3.4] Berg, S.V. – Tschirhart, J.: Natural Monopoly Regulation. Cambridge University Press 1988

[8.3.5] Bernstein, J. I. – Sappington, D. E. M.: Setting the X Factor in Price Cap Regulation Plans. Journal of Regulatory Economics, Vol. 16 (1999): 5-25

[8.3.6] Cooper, M.: Universal Service. A Historical Perspective and Policies for the Twenty-First Century, *http//www.benton.org/Library/Prospects*

[8.3.7] Crandall, R. W. - Waverman, L.: Who pays for Universal Service? When telephone subsidies become transparent. Brookings Institution Press 2000

[8.3.8] Gasman, L.: Universal Service: The New Telecommunications Entitlements and Taxes, Cato Policy Analysis, No. 310 1998

[6] Avoidable cost of service is the cost, which the operator had not have to bear if there would be no universal service obligation imposed on it.

[7] There are many types of benefits from ubiquity, goodwill, customer information, etc.

[8.3.9] Intven, H. (ed.): Telecommunications Regulation Handbook, InfoDev The World Bank, 2000

[8.3.10] Kahn, A. E.: Letting Go: Deregulating the Process of Deregulation. The Institute of Public Utilities and Network Industries, 1998

[8.3.11] Laffont, J.-J. - Tirole, J.: Competition in Telecommunications. MIT Press, 2000

[8.3.12] Mitchell, B.M. – Vogelsang, I.: Telecommunications Pricing. Cambridge University Press, 1991

[8.3.13] Mueller, M. L. Jr.: Universal Service. MIT & AEI Press, 1997

[8.3.14] Neu, W. – Stumpf, U. – Nett, L.– Schmidt, F.: Costing and Financing Universal Service Obligations in a Competitive Telecommunications Environment in the European Union, WIK, 1997

[8.3.15] Sidak, J. G. - Spulber, D. F.: Deregulatory Takings and the Regulatory Contract. Cambridge University Press, 1997

8.4. Technical Regulation

Sándor Szilágyi, dr., author

György Takács, PhD, reviewer

The general necessity of the regulation is based upon the development of the market in the expected direction. Within that function, one of the important areas of the regulation is to handle with the technical aspects which have significant influence on the networks and the services.

The first area where an active regulation is necessary is the exploitation of the scarce resources, which have considerable impact on the telecommunications activity and therefore they should be managed reasonably in the interest of the whole society. The questions of management of numbers, names, addresses and moreover, the management of frequency spectrum will be treated in chapters 8.6 and 8.7.

It is a well-known fact that the instruments of the technical "soft" regulation, the standards are forming the basis of the consistency of technical solutions, since the standards are accepted on the strength of consensus or, at least, on a qualified majority. Although the standards are not obligatory, but theoretically the service providers have the willingness to accept them in order to cooperate with each other, and to realize their investments by purchasing standardized equipment on the market at favourable prices. Are there any cases when the standards are not enough to realize technical compatibility? In the present chapter, the role of standardization in the regulation will be examined.

The next question is the following: do technical questions have legitimacy in those cases when the regulation intervenes in the activity of the telecommunications undertakings by legal and/or economic means? Generally, the answer is yes since the legal and economical regulation concerns telecommunications systems, networks and services with their attributes and parameters. Another example of that is also given in the chapter 8.5 demonstrating the most critical issue of the regulation: the interconnection of networks. Another answer is presented by the control of service quality, which is practised by the authority in the framework of market surveillance?

Further on, we will mention the technical fundamentals needed for the regulatory interventions and the instruments of the regulation in this respect.

8.4.1. Instruments of the technical regulation

The technical regulation is based upon the recommendations of International Telecommunications Union (ITU), the decisions of World Radio Conference (WRC) and European Communications Committee (ECC) and the directives and regulations of EU as well as upon the standards. By standards we mean the national (Hungarian), European (ETSI, CENELEC) and international (ISO, IEC) standards as well as some *de facto* standards, e.g. RFCs of the Internet Engineering Task Force (IETF). The technical regulation may also consider several information documents (e.g. Communications of the European Council, documents ETR, TR, TS, EG of ETSI, ITU Handbooks, EOQ papers etc.).

Essentially, technical regulation is compiled of a set of statutes and the guidelines of the authority. The statutes are more important since the guidelines may not contain any mandatory requirements, rather they give directives for certain questions of customers raised during the procedures at the authority. The technical statutes are typically Ministerial decrees; however, if their significance outgrows the sector they can be issued in the form of Governmental decree. A characteristic example is the Governmental decree on the National Table of Frequency Allocation [8.4.1] or that on management of telecommunications numbers and addresses [8.4.2]. Those technical regulations related directly to civic rights are defined by the Communications Act itself. A Ministerial decree may state the mandatory application of certain standards stipulating measures for protecting personal health, safety, security or extraordinary values, detailed and precise description of which can be found in Hungarian standards which do not give space for many alternatives by their own nature.

Fundamental Technical Plans

In certain critical period of the development of the telecommunications, the Fundamental Technical Plans have outstanding significance which define the attributes of some network types in order to ensure their interoperability and their

dependability on the expected level. Such characteristic Fundamental Technical Plans e.g.:

- Structure and Traffic Routing Plan of the public telephone network,
- Numbering Plan of the public telecommunications networks,
- Connection Plan of the public telecommunications networks which describes the subscriber's and network interfaces,
- Transmission Plan of the telecommunications networks,
- Signalling Plan of the telecommunications networks,
- Dependability Plan of the telecommunications networks,
- Security Plan of the telecommunications networks and services etc.

Some of them have been issued in the form of Ministerial decrees and have had special impact on the course of the privatization of the sector. At that time, the originally single uniform PSTN had been divided into networks of different service providers, moreover, with the introduction of the GSM mibiée radiotelephone and ERMES paging networks a number of new network types had also appeared to be connected seamlessly into the telecommunications blood circulation of the country.

The European Directives are also stressing that the regulation should regard to the integrity and secutity of the telecommunications networks. According to an EU document they are considered as follows:

- integrity is the maintenance of the operation and interworkig of the networks and services either in normal circumstances or in case of traffic overload or at an unexpected failure of the essential network elements;
- security is the protection of the network and service provision in normal circumstances as well as in emergency situation, during disaster, terrorist actions or wartime.

8.4.2. Function of the technical standards

Although the previous chapters are not focused to the broad assortiment of the telecommunications equipment, reference has been made to the standards in the cases of both the networks and the services. Now we point out an important kind of equipment standardization of which is of paramount importance: the telecommunications terminal equipment, it be either radio or wireline devices. When a terminal equipment complies with the specification of a certain service provider or network operator only, but it may not be attached to the network of a competitor

without disturbances or difficulties, a niche market is going to be formed, the equality in competition is injured and the suppliers cannot reach the desirable level of economy of scale. The user's complaints are hardly evaluable since it is not expected that they were so familiar with the technical details differentiating the networks of the service providers. One of the results of this situation is that the users can purchase the terminals exclusively in the retail shops controlled by the service provider where the availability of the full assortiment of terminals, including the most modern ones, is not guaranteed. Such segregation of the marketplace is not only restricting the choice but also has a price inflating effect, owing to the peculiar alliance between the service provider and the supplier.

Standardization of terminal and radio equipment

Formerly, the European Union wanted to arrange the standardization in such manner that

- it formulated the so called essential requirements [8.4.3],
- ETSI was charged to collect from its standards the requirements which are essential and together with the measurement methods, to harmonize them with Member States, and to publish them under the title "Technical Basis of Regulation" (TBR); then, having finished this task,
- incorporate the TBRs as "Common Technical Regulations" (CTRs) one by one into the legal documents for the Member States.

The standardization work in ETSI resulted in approx. 38 TBRs which include almost the whole assortiment of the telecommunications terminal equipment. Hungary has homologized them in the framework of access preparation to the EU (see Hungarian standards No. MSZ 25001 and consecutives) and applied them in the procedure of the national type approval.

No matter how proper had been this method in sake of establishing a unified terminal equipment market, however, a continuous lag appeared behind the technical development. Therefore the EU has made the conditions lighter for entering the marketplace by ceasing the mandatory force of CTRs and, in a new Directive called R&TTE (Radio and Telecommunications Equipment) [8.4.4] introduced an "*ex-post*" regulation instead of the former "*ex-ante*" one. That means that the legal precondition of bringing the equipment into the market is the compliance verification instead of type certificate of the Authority. Every supplier has his own risk that his equipment

may be banned from the market in the process of market surveillance if it turns out that the terminal equipment does not fulfill the essential requirements re-formulated in the new Directive. Consequently, there is no obligatory requirement any more; the suppliers may test the compliance upon requirements in a way differing from that in the standards. In so far as the ETSI standards are harmonized, that is, they are accepted by all Member States, are forming a possible, but not exclusive basis of the compliance verification.

It can occur that a certain terminal equipment is manufactured in accordance with a supplier's specification which includes the testing of fulfilment of all essential requirements, thus, it is put into the market in a justified way. This case is expected to happen with the brand new equipment and its rapid bringing into circulation for sake of the extraprofit would be not hampered by the well-known slow and clumsy procedure of standardization.

The strongest areas of the standardization are, so far, the health protection of the users and the electromagnetic compatibility. The harmonized standards issued up to the present are related almost exclusively to the above mentioned two areas. The issue of electric safety is also covered by an EU Directive [8.4.5] introducted in the Hungarian legislation too [8.4.6]. New elements of the health hazard are the physiological influence of the electromagnetic radiation and the effect of laser radiation. The homologization of the EMC Directive [8.4.7] has been also carried out [8.4.8].

The standards for radio non-terminal equipment are being also elaborated in ETSI. However, at the time the draft of this book was ready, the EU regulation did not made tangible results by issuing harmonized standards for those essential (as the Hungarian legislation puts it, "fundamental") requirements which are related to

- interworking with the network and with other terminals through the network,
- preventing harm to the network and misuse of the network resources,
- protection of personal data and rights of the user,
- access to the emergency services,
- use of services by handicapped persons.

Standardization of the interfaces

The above mentioned R&TTE Directive which is homologized by Hungary too [8.4.9], has effects not only for the terminal equipment. The Directive stipulates that the network operators or service providers should notify their interfaces provided for attachment of the terminals to their network. Formerly the interfaces had been defined, although indirectly, by the CTRs, but recently it became an issue of the declaration of the service provider. The National Regulatory Authority is mandated to register and to publish them, providing information for terminal equipment suppliers in order to consider the specification of the networks when the suppliers design the terminals.

This change apparently eliminates the uniformity of the networks providing similar basic services. However, the practical uniformity of such networks has never been achieved in the EU. The simplest examples are the telephone networks of the Member States which differ, at least, in the form and connection scheme of the wall socket. Therefore the new Directive gives the suppliers an opportunity to unify their terminal equipment in such a way that their uniform realization does not exclude the matching with the traditional differences of the interfaces in the different countries. Apart from the physical attachment, modern terminals are to be software programmed in such a way that they can fulfil the requirements in a broad assortiment of networks.

A characteristic example of these flexible terminals is the so called software radio equipment, that is the terminals which download the programme needed for interworking with the network from their home network; thus the interface specified in advance as above serves only for the first attachment and for the interworking during the first download. Expectedly, the future mobile broadband radiocommunications system, the UMTS (Universal Mobile Telecommunications System) will have such kind of terminals which will be capable as for telephony so for multimedia-type telecommunications too.

The R&TTE Directive cited above prescribes also that the European Commission defines the equivalence of the interfaces and publishes it in the Official Journal of EU. Until Hungary will join the EU, this task is mandated to the National

Authority by a Hungarian decree. Hence it is a typical technical function of the Authority.

Another area of the interface standardization is that of the interconnection interfaces, i.e. those connection surfaces through which the networks with different ownership, different functions but with a uniform system of message routing can be interconnected in order to reach the subscribers i.e. the services of each other. The role of the regulation is to require that a part of the interconnection and network access contracts be the technical specification of the networks in order to prevent the debates with the necessity of the intervention of Authority.

Standardization of networks, network equipment

There is a strong, worldwide tendency for description of the networks by their interfaces towards the outer world and not bothering with the devices used for realization of the services provided through the interfaces ("black box"). Of course, the services are considered together with their quality. The Hungarian regulation omitted already at the start the transmission, power supply, network management and cable TV equipment from the set of equipment submitted to an obligatory type approval [8.4.10], but for the switching equipment the licence for commercial circulation and for putting into operation was mandatory. This condition is expected to cease and the Authority is charged for checking the fulfilment of, at least, two requirements: health protection of the users and compliance with the EMC reqirements, in the framework of market surveillance. Although the EU does not have this issue in terms of reference to elaborate harmonized, equipment-specific standards yet, the existing general requirements can be applied here too.

Relating to the user's health protection, there is a Hungarian decree [8.4.11] which makes a set of standards of safety character, e.g. protection against the athmospheric effects, effects of power lines, obligatory. This set of standards should be extended in the future, e.g. by those related to safety of devices using laser radiaton (see Hungarian standard MSZ EN 60825-1:2000), by standards concerning issues of the health protection against the electromagnetic radiation which issue is already dealed with in a Hungarian decree [8.4.12] and so on.

Standardization of services

Standardization of services provided through the networks can be realized in two areas: for basic services and for supplementary services. The value added services are so far-reaching due to their included informatics and content services (see Chapter 5.2) that their standardization is not reasonable. Typical areas of the standardization of basic and supplementary services are the services provided on networks built in accordance with open, international standards (e.g. GSM, ISDN, UMTS). An essential part of these standards is the service description. The regulation is confined to supervise the real operation in the interest of proper national and international interworking of the networks.

8.4.3. Regulation of the quality of telecommunications services

A separated area of standardization of the services is the quality of services (QoS). There are international recommendations, handbooks which describe the characteristics of QoS, measurement methods of its parameters enabling a comparison between the service parameters of telecommunications companies in a competitive environment. Since the level of QoS - excepting the quality of universal service - is an issue of declaration of the company, the competiton can be fair if the service providers declare their QoS in the Code of Practice with conform principles, in a comparable manner. The regulation should define the set of quality parameters and their measurement methods which should be stipulated and eventually the benchmark values declared in their Code of Practice.

Additionally, the regulation is aimed at providing the service providers with certain minimum values which oblige them to fulfil certain quality expectations independently of the price/quality ratio, in the user's interest. This means that the regulation should prohibit the provision of an impracticable service even for a very low price.

Special stress is given by the related EU Directives to the quality expectations on the field of the so called universal services. The universal telecom-munications service can be used by the inhabitants independently from their geographical locations, at an affordable price and the service provider is obliged to enter a

subscriber contract with any demanding customer. Today, this universal service is provided by the concession telephone companies.

There are some quality requirements in the concession contracts even now which will be needed in the future until the user cannot freely choose his service provider. The universal sevice provider can sell and maintain the access line at a lower price if his costs are reimbursed by the state in the framework of the contract. On the other hand, the customer has no choice; the service provider can misuse the situation and can provide rubbish services without any minimum quality treshold.

The quality parameters of the telecommunications service can be defined, even in most cases, measured. The EU Directive on open network provision (ONP) [8.4.13] recommends the following quality parameters to be defined by the National Regulatory Authorities:

- supply time for initial network connection,
- fault rate per connection,
- fault repair time,
- call failure rate,
- dial tone delay,
- call set up delay,
- statistical parameters of the transmission quality,
- response times of operator's services,
- the proportion of operability of payphones and cardphones,
- accuracy of billing.

These are statistical parameters and they are not peceivable directly by the user. They can be evaluated by the Authority relying upon the statistics of the service provider. Both the minimum treshold values and the measurement methods are to be regulated in order to comply with the Authority's obligations in the mentioned Directive. Similar tresholds and methods can be formulated for other telecommunications services, too.

The quality attributes perceivable directly by the user can be figured out by polling (e.g. via questionnaires) among the users. The user's direct experiences are generally of "yes/not" character. He surely can complain only in the case when there is no service at all, or the calls are regularly errored etc. Most of them are also typically technical parameters such as those of statistical character.

8.4.4. Other technical regulatory issues

One of other issues related to technical regulation is the regulation of tariffs. The subject is not the measure of fees but the way of their enforcement. The technical attributes of the telecommunications networks play a significant role how the service provider can form his tariff structure. One of examples was a recent amendment of tariffs, when some service providers switched over to the second-based billing but they wanted to stop the transmission of 12 kHz tariff pulses to the subscriber's line. The technical regulation had to intervene to prevent this stopping since formerly the service provider had offered to send those pulses on demand, in order to operate the subscriber's remote charge meter if he had purchased this device. Ceasing this service would hurt the interests not only of these subscribers but also the interests of numerous PABX operators who have used the service for accounting the traffic costs of individual extensions for the calls sent into the public telephone network. A significant surplus cost would occur at these PABX operators, therefore the regulator obliged the service providers to continue the provision of this service.

When the issue of tariff sharing and accounting was dealt with during the interconnection negotiations, one of the decisive questions was the capability of the controlling processors of the telephone exchanges for the storage of different charges according to primary areas and hours of the day. The regulation had appeared here in the form of seeking a technical consensus and a harmony of technological opportunities.

In the course of the development of the telecommunications certain differences appeared among the countries which differences have been included in the standards as national options. An outstanding example is the European standard ETS 300 001 which is actually an encyclopaedia of specifications of the telephone sets in the European countries. Both in the ITU-T Recommendations and in the European standards there are national options to be defined by the National Administration. Typically, the standards describing signalling systems are of that kind. In these cases, in order to save the network integrity within the country, it is the task of National Regulatory Authority - relying upon a consensus with the service

providers and the suppliers - to define those technical parameters which will then become national requirements.

8.4.5. Following the technical development

The technical regulation is repeatedly meeting the challenge of continuous development of the technology which is reflected not only in the sophistication of operational protocols and controlling programmes. More important are the fundamental changes which demand a paradigm shift on the field of telecommunications.

Formerly a latent but at the end of 90-th years openly appearing tendency is the dynamic spreading of the use of Internet protocol (IP). The bearers are those new applications previously not heard of in the society, such as electronic mail, WEB-browsing, electronic commerce and many others.

One of the application of the IP protocol which has shattered fundamentally the telecommunications world is the IP telephony, the Voice over Internet Protocol (VoIP) technology. The technical regulation had and has to deal with this issue in such a way that it has to consider the legislation in force, the existence of concession contracts, but not to hinder the technological development by banning the applications.

Until the solution of concession monopolies, the only way to allow national and international long dictance calls using VoIP technology evading the monopolistic service provider assigned for this function, has been to withdraw this technology from the services of real-time telephony while the regulation prescribed some artificial quality-spoiling parameters for it. The characteristics of these parameters are also confirmed in a EU Communication [8.4.14].

The telephone service realised by using VoIP technology obviously brings along the possibility of - jittered - delay of several hundreds of milliseconds as well as that of interruptions owing to packet losses. Utilizing this, the Regulatory Authority required a delay of minimum 250 ms and that the guaranteed loss of packages should be not less than 1 %. From the point of view of speech quality, these parameters have made it differentiable from the classic, circuit-switched technology.

After having the monopolies fully solved, a great task falls upon the regulator, because - in the interests of the users, i.e. of the society, the VoIP service providers should be further stimulated on enhancing their QoS so that it would be not worse, but better than an acceptable treshold value.

Next step of the development of the technology, made by the industry, is the use of the IP protocol or, at least, the packet switching, in the radiocommunications networks.

What is the dilemma?

The duplex speech transmission called telephony is appearing nowadays in many habits with similarities but also with some differences.

Technology	wire (cable, fiber)	radiowaves (p-p, p-mp)
Circuit-switched	wireline telephony (PSTN/ISDN)	mobile radiotelephony (GSM) cordless telephony (CT)
Packet-switched	IP telephony (Internet)	IP radiotelephony (GPRS. UMTS)

When the QoS is compared, the squares are characterised by different parameters:

Quality parameter	wireline telephony	mobile radio-telephony, CT	IP telephony	packet-switched radiotelephony
noise (qdu)	1	3.5…7.5*	?*	?*
delay (ms)	~0 (<400)	>100	>100	>100
drop-out	0	due to propagation	due to loss of packets	due to propa-gation and loss of packets
flow control necessity	not necessary	due to reflections	due to different paths	due to reflections and different paths
call routing	SS#7	SS#7	IP/SS#7	IP/SS#7
24 h operability	with remote feeding	with battery	with unbreak-able power supply	with battery
Lifeline operation	yes	not typical	not resolved yet, but necessary	not typical

* value of the qdu is a function of speech coding procedure applied and is usually the more the less the resultant data speed is.

The technical regulation must face to this dilemma which can be formulated in the following way: the QoS should be regulated in such a way that considering the allowances made by the users against the quality for the sake of

- on the one hand, the mobility,
- an the other, the opportunity of the multimedia services are acceptable in what degree for the society and where can be found the treshold of unacceptable

service reflected in the parameters? There is no doubt that the international telecommunications organizations and the pan-european regulation will give some kind of answer on these questions.

References

[8.4.1] Governmental decree No 221/1999. (XII.20.) on the National Frequency Allocations Plan (NFAP)

[8.4.2] Governmental decree No 75/2000. (V.31.) on the management of telecommunications numbers and addresses and on the rules of procedures of that

[8.4.3] Directive 98/13/EC of the European Parliament and of the Council of 12 February 1998 relating to telecommunications terminal equipment and satellite earth station equipment, including the mutual recognition of their conformity

[8.4.4] Directive 1999/5/EC of the European Parliament and of the Council of 9 March 2000 on radio equipment and telecommunications terminal equipment and the mutual recognition of their conformity

[8.4.5] Council Directive 73/23/EEC of 19 February 1973 on the harmonization of the laws of Member States relating to electrical equipment designed for use within certain voltage limits

[8.4.6] Ministerial decree No 79/1997.(XII.31.) IKIM on safety requirements of certain electrical products and on the conformity assesment of those products

[8.4.7] Council Directive 89/336/EEC of 3 May 1989 of the Member States relating to electromagnetic compatibility

[8.4.8] Joint Ministerial decree No 31/1999. (VI.11.) GM-KHVM on the electromagnetic compatibility

[8.4.9] Ministerial decree No 3/2001.(I.31.) MeHVM on radio equipment and telecommunications terminal equipment and the recognition of their conformity

[8.4.10] Ministerial decree No 8/1991. (III.14.) KHVM on the fees paid for postal and telecommunications procedures of the Authority

[8.4.11] Ministerial decree No 27/1994. (IX.29.) KHVM on declaring certain transports, communications and water managements national standards as obligatory

[8.4.12] Ministerial decree No 32/2000. (XI.16.) EüM on the health risk treshold values of electromagnetic radiation emitted by radiocommunications establishments

[8.4.13] Directive 95/62/EC of the European Parliament and of the Council of 13 December 1995 on the application of open network provisio (ONP) to voice telephony

[8.4.14] Communication 2000/C 369/03 from the Commission: Status of voice on the Internet under Community law, and in particular, under Directive 90/388/EEC

8.5. Interconnection and Unbundling of Networks

Sándor Szilágyi, dr., author

György Takács, PhD, reviewer

The Internet service providers are frequently boasting that Internet is "the network of networks" which makes access to the services presented at any point of the world. This formulation is also true for fairly older networks, e.g. for the worldwide telephone network, since the national telephone networks were already interconnected in the first half of the 20th century in such a way that the calling party was able to get connected - in the beginning, via manual international operators, later on, using long distance dialling - to any fixed or mobile subscriber of the world. This has been facilitated by appropriate public demand as well as by matching the different existing speech transmission specifications. Thus, a commonly accepted system of signal levels and that of addressing of the calls has been established so that the calls could be conveyed from one network to another. Besides the interconnection, other forms to connecting networks together has been implemented such as network access and unbundling of the local loop.

It is the prime interest of the society that in a competitive environment, when several networks participate in the conveyance of the messages, the users should not suffer disadvantages in this case, compared with the case of the conveyance of calls in a single, monolithic network (see Chapter 8.3.2). Therefore, the regulation should monitor the internetwork contracts signed by the network operators or service providers. The set of commercial links established by the internetwork contracts may give the service providers an opportunity for a discriminative or exclusive behavior at the competition.

8.5.1. Necessity and Conditions of the Interconnection

During the development of telecommunications technology, in one of the archetyps of telecommunications switched networks, in the telephone network, early demand had emerged for an opportunity to communicate with the subscribers of other telephone exchanges and vice versa. Later on, this demand had been

expanded to the accessibility of the subscribers of larger territories such as other countries. That development has brought along the structuring of the networks; the local networks incorporating the local access of the subscribers; the transit networks intended for interconnecting the local networks and an international network have been formed. International gateway exchanges as well as a whole system of international links have been implemented.

In the course of marketization of the telecommunications, demand emerged not only for interconnection of the networks of other territories but also for the interconnection of the networks of parallel, competitive operators providing services within the same territory. Similarly, a number of transit service providers – carriers – have appeared, each interconnecting the local networks through their own transit network. Consequently, a marketplace has been formed, not only between the service provider and the subscribers but also among the service providers.

During the past 100 years, interconnection has taken place – apart from the public networks – between business, corporate and private networks, including even networks of different purposes (e.g. between telephone and data networks), too. Therefore, it has became inevitable to formulate the essential attributes of the interconnection.

This formulation says that the interconnection is a connection between telecommunications networks which enables the users of one network to exchange information with the users of the other network as well as to reach the services of the other network, which are provided by the service provider of the latter network or by third parties who can be reached through the latter network. Consequently, the networks to be interconnected should be technically compatible. Further on, it is essential that the legal relation of the users (subscribers) attached to any of the networks to their service providers remains unchanged after the interconnection, in spite of the fact, that they are using the services of another network operator. With this approach, the interconnection is conceived as a symmetric interworking system of networks.

In the Interconnection Directive [8.5.1] the European Commission extended the definition given above to the case when a single service provider uses more than one network. The Directive contains the following definition:

'interconnection' means the physical and logical linking of telecommunicatons networks used by the same or a different organization in order to allow the users of one organization to communicate with users of same or another organization, or to access services provided by another organization. Services may be provided by the parties involved or other parties who have access to the network.

It is obvious that there is a significant attribute of the interconnection: the reciprocity, since it extends the services equally to users of both networks without any change in their legal relationship with the original service provider.

In Hungary, the networks have developed within a monopolistic environment in such a way that no differentiation described above could really be formed in the public telephone network. The single public telephone network of the country operated by the state owned company MATÁV Co., which has been a multi-level hierarchical structure with sectorial, supernode and node areas, but neither their number nor their system had been in accordance with the perspective digital swithing technology and with the changing cost ratio of transmission and switching. In the ’80s, after a detailed techical-economical investigation a new structure was defined for the future digital network, the number of network levels and of their exchanges. On the lowest level 54 areas were formed. In 1993 a Ministerial decree [8.5.2] was issued which stipulated the territory of local telephone areas - the so called *primary areas* – as a minimum unit of the network to be privatized, and which defined the national long distance (transit) network which would interconnect the primary areas with each other (this concept has been mixed up incorrectly with the concept of national backbone network). That means that the first "unbundling" of the single network, i.e. the first technical-economical separation was carried out.

However, this separation has been complete only for the new private local telephone concession companies (see the details in Chapter 5.2.). The MATÁV Co., although being granted a separate concession for its local telephony services and another separated concession for the services of the national and international long distance network, the two kinds of services could not be divided due to the particular monolithic architecture of the company and later on, the Minister merged the two concession contracts.

In the scenario described above, all other service providers - including the mobile radiotelephone companies - have had to interconnect their networks with that

of MATÁV Co., the dominant telephone service provider. An exclusive right of the latter for national and international long distance connections has been stipulated in the contract has meant that no direct international access has been available even for the mobile radiotelephone users. Thus, for example, the traffic of international mobile roaming has been enforced to be conveyed through that single transit network, in spite of the fact, that it was hardly a business of the fixed line telephone operator.

In the developed countries, especially within the EU, the introduction of the carrier selection is in the agenda of the network operators (see also in Chapter 8.6.). Since in those countries there are specialised interconnecting transit networks and/or service providers who are competing with each other, right is granted to the subscriber to select a carrier for his/her national or international long distance calls. The subscriber's right mentioned above should be provided partly in the way of pre-selection (when the subscriber enters a contract with the national and the international long distance carrier). In addition, the subscriber should have an opportunity to choose a carrier independently of his/her pre-selection, on a call-by-call basis.

A precondition of the use of carrier selection facilities is that the given local operator has to be interconnected with the selectable carrier. In the European Communities, the local telephone operators are obliged to get interconnected with the carrier in case when they receive an interconnection offer from the long distance operator.

There is a frequently asked question: why the Internet service providers operating on their "network of networks" do not enter any interconnection agreement between each other? There are some historical reasons. Originally, the Internet had been formed as an infrastructure of non-profit civil undertakings such as high schools, libraries etc. When interworking of two local area computer networks (LANs) was needed, the interested parties jointly rented the transmission paths, hired and maintained the necessary routers and other equipment. Tariffs for conveying the data and, consequently, tariff sharing have been not necessary. Notwithstanding that the parties certainly enter an agreement on bearing the maintenance costs, on handling the fault detection and repairs, such agreement does not include any commercial element and therefore it is not called interconnection contract.

Interconnection payments are a separate economical issue. National Regulatory Authorities regulate them in different, country-specific ways. In some countries the interconnection fee is used to compensate (partly or wholly) the losses of local network operators due to the universal service obligation. According to the EC's recent standpoint, the interconnection fee should be proportional to the long-range incremental costs.

Although it is not a pure technical issue, but the definition of the interconnection marketplace slightly touches the topic of technical regulation. The EU legislation takes primarily the termination fee, i.e. the fee paid for the handling and termination of the calls originated in other (in case of telephony, local or radiotelephone) networks, into account. Consequently, the major players of the interconnection market are to be found among the service providers with significant market power either on the fixed telephone or on the mobile radiotelephone market, instead of the players of the carrier market.

According to the Directive of EU on the subject [8.5.1], the service providers qualified by National Regulatory Authority as those having Significant Market Power (SMP) on the interconnection marketplace, are submitted to an obligation to issue on demand an interconnection offer related to a network point which does not coincide with the network termination point offered usually for end-users.

According to the related EC Directive [8.5.1] the service providers identified by NRA as having significant market power (SMP) on the interconnection market are obliged to present an interconnection offer on demand for a network termination point which differs from those offered usually for interconnection.

8.5.2. The interconnection contracts

Legal and commercial background of the network interconnection is the interconnection contract. Its essential requirements defining the content elements are [8.5.1]:

a) maintaining of operability in case of catastrophe (extraordinary wheather, earthquake, flooding, outbreak of fire);

b) maintaining the integrity of the network in case of overload, technical failure, perturbances of power supply;

c) interworking in the service provision with an adequate quality;

d) maintaining the data protection and secrecy.

Requirement a) means an obligation to cooperate in case of failure or drop-out of physical devices of interconnection when cooperation and joint measures are necessary. The contract should stipulate these cases and should indicate the high management level where the necessary joint steps can be made for restoration of the interworking capabilities of the network.

Requirement b) reflects the responsibility of the parties in the area of cooperation in normal network operation, during fault detection and repair. The requirement of integrity prescribes cooperation not only in the maintenance of the telecommunications basic networks operability but also in the network management.

Requirement c) contains the contractual statement and the implementation of accessibility of services for their subscribers in the other party's network. If the mutual provision of a certain service meets difficulties, the contract should stipulate it.

According to Requirement d), the parties should keep the privacy of subscriber's messages of each party. To this requirement refers the confidential handling of user's data provided each other by service providers in the course of complain conduct, fault elimination etc.

Beyond the essential requirements, the contract should also stipulate certain technical issues, namely the description of internetwork interfaces. In a modern circuit-switched network it means, at least, four interfaces:

- interface of speech circuits,
- interface of the signalling network (e.g. SS#7),
- interface of the network management system and
- synchronisation interface.

In a packet switched network the transmission path of the information payload and that of the signalling are not necessarily divided. Some networks are using decentralised synchronisation (e.g. based on GPS); thus they don't need any kind of synchronisation interface.

The EU Directives recommend the use of standardized interfaces specified by National Regulatory Authority.

It is expected that one of most difficult and debatable part of the interconnection contract might be the chapter of interconnection fee. The quoted Directive [8.5.1] prescribes that the prices of the service provider with significant market power should be transparent and cost-oriented. The service provider is obliged to publish the so called Reference Interconnection Offer – approved by the NRA – containing the main components of cost calculation which forms the basis of tariffs.

In addition, the regulator should require that the service provider with significant market power act without discrimination at entering the interconnection contract, provide all necessary data to the partner. The contract shall be presented to the NRA which is mandated to show everybody all the parts of it excluding parts representing business secrets.

8.5.3. Network access

The networks may be connected together not only in the above described symmetric manner. There is another case when a network is connected to another in order to use the latter or one part thereof for the services of the first network operator. The simplest form of this case the leased line service (see in Chapter 5.6) but now we are studying mainly the more sophisticated forms of network access.

The simplest and long used service of that kind is the PABX dialling-in. When the calling numbers of the PABX extensions are public and they are included in the telephone directory, the public service provider range them into the subscriber capacity. Thus, the service provider can enlarge his number of subscribers while saving the investment of the subscriber's switching stage as well as that of the local subscriber's loops, he uses the private network of the PABX for terminating the calls originated within the public network. In this case the operator of the PABX provides network access service for the public telephone service provider.

There is another related issue: today the public service provider does not pay to the PABX operator for this opportunity of excess traffic capability, what is an example of relics of the monopolistic era when the public provider endorses his market power.

A possible case of the network access when the Internet service provider undertakes the costs of the public telephone subscriber's access in such a way that he "hires" the telephone network, i.e. reimburses the "dial-up" access costs for the telephone operator.

A case of the network access is also when - in lack of wireline connection - the telephone service provider entrusts the mobile radiotelephone operator with the provision of services for certain subscribers. This case exists in our country, since, for the sake of deliberation of the 900 MHz band, the original radio local loops were eliminated and the telephone service provider requested at one of the GSM operators for the temporary provision of services to the subscribers concerned. However, the fixed telephone subscriber does not become a mobile one; the service provider installs in the residence or house of the subscriber a radio transceiver box which has a radio side with GSM handset capabilities, and a subscriber's side providing a two-wire analogue connection socket.

Network access takes place in the so called virtual mobile radiotelephone network operation which has numerous examples in Western Europe. If there is free capacity in a certain mobile radiotelephone network (e.g. the traffic load of the cells remains below the expectations), this free capacity may be sold by the mobile operator to a service provider who has no network of his own because he has not participated at the tender for the frequency or has not succeeded to get licence at the same place. It is possible to make such an agreement with conditions allowing a "win-win" situation: with a faster recoverage if the investment for the network operator and with an opportunity to get subscribers faster for the new service provider and to start the service before implementing his own infrastructure.

The EU obliges the service providers with significant market power to provide the network access service.

8.5.4. Unbundling of the local loop

There are two ways of unbundling the local loop: the full unbundling and the partial unbundling.

Full unbundling of the local loop

The fixed public telephone networks have a bottleneck for the competition which can hinder or retard it very much. It is obvious that the construction costs of the subscriber's loop makes at least one quarter of the investment costs of the telephone network. For the new entrants of the market it would be a large and immediate expenditure to lay down a second loop parallel to the existing loop when the subscriber wants to switch over to this new service provider; notwithstanding that it would be deeply unreasonable. Therefore the regulation should oblige the incumbent SMP operator to surrender the metallic local loop of his subscriber to the new entrant for use, if the subscriber has gone over to the latter.

Hence it is a leased line contract with two differences apart from an ordinary analogue leased line contract:

- The original operator does not any transmission of signals on the unbundled section of the line, thus it is a simple leasing contract of a passive copper pair,
- when defining the tariff, those costs should be separated which are not directly related to the maintenance of the loop concerned ("unbundling"),
- the incumbert operator should be obliged to let the equipment or the devices of the new entrant to be placed at the switching side end of the loop (common placement, "collocation").

The EU attaches primary importance to the issue so that the local loop unbundling obligation has been published in the form of regulation [8.5.3] which – differently from the directive – is a EU statute to be implemented without homologization in the Member States. However, this regulation is considered as temporary and valid until the new service provider establishes his own infrastructure.

This solution, of course, does not exclude other solutions for the subscriber access (e.g. point-multipoint radiocommunications systems, cable TV networks), but these are not always competitive with the metallic loop.

Partial unbundling of the local loop

The technological development has made it possible to realise the transmission of further messages along the metallic local loop, beyond the usually utilized bandwidth (see Chapter 2.10). The services available through this transmission capability generally are not, however, telephone services. Although the

first investigations of better utilization of the loops had been focused to transmission of TV programs, later on the Internet service providers used the technical opportunity.

This kind of unbundling is similar, to a certain extent, to the full unbundling of the local loop, since the incumbent operator is obliged for this service too due to the fact that this service may have positive effect on the provision of wideband data transmission, preferably of the Internet.

As a matter of fact, it’s all about a special kind of line leasing while a contract is entered not on the full usage of the physical line but on the usage of its defined frequency band only. The leased line contract has the same attributes as those mentioned in the previous chapter. The difference is that the broadband transmission of the signals always requires the installation of special equipment (data multiplexers, splitting filters) which can be handled from the point of view of the collocation in a more complicated way than in the case of full unbundling, when handling of the simple pair of wires occurs.

Reference offer for the local loop unbundling

The abovementioned regulation of the EU requires that the SMP service provider owing the subscriber's loop has to publish his reference offer which should stipulate (in a shortened way)

- the network elements and sites providing physical access as well as the technical conditions of the access,
- technical data of the twisted copper pair,
- ordering and installation procedures and restrictions,
- data of premises with the above access points from the point of common usage,
- security and visiting requirements and constraints of access to the commonly used premises,
- the conditions of access to managing, information systems and databases of the incumbent,
- the time schedule and price of provision of the ordered access,
- the conditions in the area of service quality, fault detection and repair works.

References

[8.5.1] Directive 97/33/EC of the European Parliament and of the Council of 30 June 1997 on interconnection in telecommunications with regard to ensuring universal service and interoperability through application of the principles of Open Network Provision (ONP), OJ No L199, pp. 32- 52 (26.7.1997

[8.5.2] Ministerial decree No 26/1993. (IX.9.) KHVM on the Structure Plan of the public telephone network

[8.5.3] Regulation (EC) No 2887/2000 of the European Parliament and of the Council of 18 December 2000 on unbundled access to the local loop, OJ No L336, pp.4-8. (30.12.2000)

8.6. Numbering

Géza Gosztony, PhD, author

György Takács, PhD, reviewer

For the user of telecommunications numbering appears mainly as a tool to indicate the destination of a message or the choice of a service. For those being more familiar with the characteristics of telecommunication networks and services it seems to be obvious that for exchanging information between human beings and/or machines it is always required to know where the message has to go to and where does it come from. Numbering (or more generally the use of identifiers) caters for the proper flow of information in and between networks. The aim of this subsection is to give a basic understanding of the various aspects of numbering in telecommunications and to include some guidelines for further studies.

8.6.1. Numbers, names, addresses - identifiers

In the telephone directory one can find the telephone number of Mr. Smith. One gets this information based on his name. The number might indicate also the address of Mr. Smith who lives in 15. High Street. However, if Mr. Smith has asked for a follow-me service the original number indicates only his name and the intelligence built into the network will find him using an other number used for routing purposes indicating his momentary address e.g. that of Ms. White. The same happens if Mr. Smith has moved to an other house but wanted his original telephone number to be ported. In addition to telephone numbers there exist also many other types of numbering resources as signalling point codes, international mobile subscriber identifiers, etc. The basic definitions below give a generalisation of the aspects mentioned valid also e.g. for machine to machine information exchange **[8.6.2]**.

- A **name** is an alphanumeric identifier used for a telecommunication service to **identify an end part** (destination) of a communication. A name is used at the service level and may be required to be portable.

- An **address** is an alphanumeric identifier used for a telecommunication service to identify and **locate an entity** in a telecommunication network. An address is used at the routing level and is not required to be portable.
- A number is a name or an address consisting of digits only.

Alphanumeric identifiers mentioned in the definitions might include numbers, letters and symbols. In English and also in EU legal texts the term numbers or numbering covers identifiers as defined above. Regulation in Hungary is using the term *identifier* as a common notion for names, addresses and numbers as defined above **[8.6.1].**

8.6.2. Numbers as scarce resources

Numbers together with frequencies are considered as being scarce resources, which require special attention. Regulation of the majority of countries and e.g. also the legislation of the European Union handles numbers in this way **[8.6.3]**. At first sight this approach does not seem valid. Strings of numbers or characters might be as long as necessary, an increase of the set of names or addresses seems to have no limits. Theoretically this is correct, however from the practical point of view there are severe limitations of expanding the range of available identifiers.

The handling of numbers (evaluation, storing, conversion etc.) requires sophisticated software and hardware of high value. The elongation of telephone subscriber numbers e.g. by one digit (ten times more capacity!) is a rather expensive action taking the addition of necessary network elements and other costs (organisation of the process, change of stored numbers in different devices as fax machines, modification of letterheads, etc. etc.) into account. The high cost related to numbering changes is one limiting factor.

Other important limitations are included in the different international numbering and addressing plans. For the sake of smooth world-wide interworking in numbering the maximum length of international numbers and codes is fixed and this limits also the length, which might be used nationally. E.g. international telephone numbers might have maximum 15 digits not taking the international prefix (being 00 almost in all countries) into account **[8.6.9]**.

8.6.3. Guidelines for regulation in numbering

Regulation of numbering is a special field for which countries with liberalized competitive telecommunications have certain obligations. The reason of this special care is the scarcity of resources and further simply the value of good, i.e. easy to remember numbers in e.g. commercial life. As a prerequisite *numbering resources should be administered by an independent body* (e.g. the independent NRA according EU guidelines). In addition (without details):

- openness to competition by *non discriminatory access* to numbering resources,
- existence of *transparent industry/administration guidelines* e.g. for numbering reservation, assignment and reclamation procedures,
- *user friendliness* of numbering schemes,
- *adequate capacity* for geographic and non geographic numbers, short codes etc.,
- *international* (e.g. European level) *harmonisation of numbering* schemes,
- *carrier selection* and *number portability* to support competition

has to be maintained **[8.6.3], [8.6.4], [8.6.5].** The new set of EU directives on electronic communications (to come into force in the first half of 2002) reinforces the guidelines above.

In many countries and almost in all European countries the assignees of numbering resources (normally telecommunication service providers and/or network operators) have to pay a fee. This fee might cover either the administrative costs of numbering only or it might contribute to the expenses of the NRA **[8.6.7]**.

8.6.4. International standardisation and guidelines

The world-wide co-operation of telecommunications networks requires guidelines for numbering. International standards are prepared and maintained by ITU-T. ETSI is responsible for European regional standards. EU regulations and directives have several numbering aspects and ETO[8] established by CEPT ECTRA has prepared many studies and guidelines for this area. (September 2001 ECTRA

[8] European Telecommunication Office

and ERC merged as ECC[9] and in a similar way ETO and ERO[10] have also merged as ECO[11]) Some important examples are mentioned below.

A short overview of ITU-T numbering activities can be found in this Section. The basic requirements for regulation of numbering together with the interpretation of and deadlines for carrier selection and number portability can be found in **[8.6.3]** and **[8.6.8].** The most important guidelines on national numbering conventions are summarized in **[8.6.2], [8.6.4], [8.6.5]** and **[8.6.6],** see Section 8.6.6. Only some basic ETSI standards are referred to here as examples **[8.6.13**], **[8.6.14].**

For different applications of international telecommunications a number of international numbering plans have been constructed by relevant bodies. These numbering plans provide the structure and functionality of the identifiers involved, give information of how to use identifiers (e.g. dialling procedures) and also cover the digit analysis required to successfully route the information (e.g. telephone calls) to be transferred. To give an idea about the areas handled a non-comprehensive list of ITU-T numbering and addressing plans is given below together with the number of the relevant Recommendation.

- International public telecommunication numbering plan –E.164, **[8.6.9]**
- B-ISDN numbering and addressing – E.191
- International Mobile Subscriber Identities (IMSIs) –E.212, **[8.6.11]**
- Plan for Telex Destination Codes – F.69
- Naming and addressing for public message handling services – F.401
- International Signaling Point Codes (ISPCs) – Q.708, **[8.6.12]**
- Data Network Identification Codes (DNICs) – X.121
- Information technology – Open Systems Interconnection – Basic Reference Model: Naming and addressing – X.650

The list above does e.g. not include those Recommendations, which handle the application of one of the numbering or addressing plans listed above. On the other hand some important internationally used identifiers as e.g. Internet names and Internet Protocol (IP) addresses are missing, since at present no ITU Recommendations exist on these topics. The reason is that except for the last some

[9] European Communication Commission

[10] European Radiocommunication Office

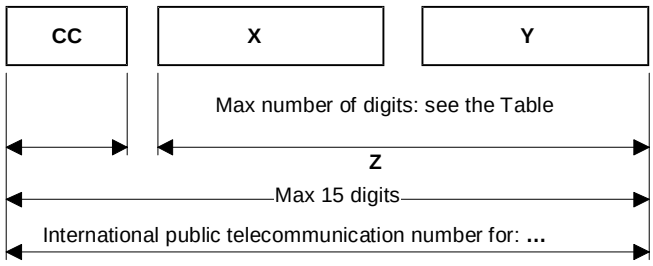
[11] European Communication Office

years the interoperability of the Internet world and conventional telecommunications was not a major issue and therefore no real attention was paid to numbering interworking i.e. interworking of identifiers used here and there. (On ongoing work see more in Section 8.6.8)

Because of its importance, as an example, the “environment” of numbering plan E.164 ***[8.6.9]*** can be found in the list below to see several aspects covered by ITU-T international standardization.

- Criteria and procedures for the reservation, assignment, and reclamation of E.164 country codes and associated Identification Codes (ICs) – E.164.1
- E.164 numbering resources used for trials – E.164.2
- Principles, criteria and procedures for the assignment and reclamation of E.164 country codes and associated identification codes for groups of countries – E.164.3
- Principles and responsibilities for the management, assignment and reclamation of E-Series international numbering resources – E.190, **[8.6.10]**
- Alternatives for carrier selection and network identification – E.164 Supplement 1.
- Number portability – E.164 Supplement 2
- Numbering plan interworking for the E.164 and X.121 numbering plans – E.166/X.122
- Application of E.164 numbering plan for
  - UPT – E.168
  - universal international freephone numbers for international freephone service – E.169.1
  - universal international premium rate numbers for the international premium rate service – E.169.2
  - universal international shared cost numbers for the international shared cost service – E.169.3
- E.164 Country Code expansion – E.193

The text of E.164 itself deals with the structure of international numbers for (i) geographic areas (ii) for ITU-T defined international global services and (iii) for international public Networks (intentionally with capital n) operated by ROAs (operating agencies recognized by an Administration). Information is included on the different parts of the number structure (see a summary in Figure 8.6.1 and



CC: Country Code **CC**: Country Code

X, Y and Z : see the Table

NOTE – National and international prefixes are not part of the international public telecommunication number

Figure 8.6.1 – International public telecommunication number structure (E.164)

associated Table 8.6.1), on dialing procedures including the use of prefixes, escape codes, on digit analysis requirements and on the application for ISDN. E.164 was and still is the most important international numbering standard.

International number for:	CC (number of digits)	X (number of digits)	Y (number of digits)	Z (max. number of digits)
Geographic areas	1 – 3	NDC – National Destination Code (national matter)	SN – Subscriber Number (national matter)	NSN – National Significant Number (15 – CC)
Global services	3	Not applicable	Not applicable	GSN – Global Subscriber Number (12)
Networks	3	IC – Identification Code (1 – 4)	SN – Subscriber Number (network matter)	– (11)

Table 8.6.1

The user of the public telephone network might indicate the international or long distance call to be initiated by the *international* or *national prefix* (being almost in all countries 00 or 0, respectively). The user might also indicate the wish to use an other numbering plan (numbering plan interworking is handled in Section 8.6.7). For international calls the CC identifies geographic area or a global service unambiguously. The CC assigned to world-wide Networks identifies a group, to find a given Network (e.g. Global Office Application network of British Telecom) the IC is also required. CCs, ICs and GSNs (see Table 8.6.1) are assigned by the ITU. On the

other hand NDCs and SNs are the responsibility of the country or the Network concerned.

The NDC part of the NSN might have different structures. It might indicate geographic areas of a country or different networks and/or services inside the country or a combination of these all. In some countries the subscriber has to dial the SN to reach others in a given geographic area, but a country may decide on a closed numbering scheme, when one has to dial always the same number of digits, i.e. the NDC is included. In this case no national prefix is required, however the NDC functionality remains. In the case of ISDN a *sub-address* might also be added which is transferred transparently to the destination designated by the NSN.

It has to be emphasized that ITU prepared and other international numbering and addressing plans take the sovereignty of national Administration explicitly into consideration. Although these plans include guidelines for the structure, etc. of nationally used parts of identifiers the final decisions referring to them is a national matter. For this reason the construction of National Numbering and Addressing Plans (NNPs) has a certain freedom but has carefully to be harmonized with e.g. regional plans.

8.6.5. National numbering conventions, situation in Hungary

National numbering conventions and the responsibilities of the NRA have been studied by ETO and quite a few recommendations can be found in relevant reports **[8.6.2], [8.6.4] [8.6.5]** and **[8.6.6]**. Regulation tasks of the NRA described in the conventions refer to:

1. The National Numbering and Addressing Plan (NNP),

2. NNP Administration i.e. *numbering policy*, dealing with: (a) the content of the NNP, (b) the structure of different type numbering resources, (c) the necessary capacity of these resources, (d) user friendliness of dialling procedures and (e) harmonisation of the NNP with international and regional requirements.

3. NNP Management consisting of (i) the practice of reservation, assignment and reclamation of numbering resources to/from service- and network providers, (ii) providing publicity and appropriate appeal procedures, (iii) regular consultation with all interested market players (iv) the supervision of the use of assigned numbering resources.

According to ETO recommendations conventions should in any case cover telephone numbers, Data Network Identification Codes (DNICs) and International Signalling Point Codes (ISPCs). In addition it is advisable to include conventions for Data Country Code ATM End System Addresses (DCC AESAs), International Mobile Subscriber Identities (IMSIs), National Signalling Point Codes (NSPCs), Individual TETRA Subscriber Identities (ITSIs) and telex numbers (this later only required where competition in telex services exists). As regards NNP administration and management issues the reader is referred to the references.

The Hungarian numbering framework

From the end of year 2001 Hungary has a liberalised telecommunication market. The basic guidelines of regulation are included in Act XL of year 2001 **[8.6.1]**. The preparation of a numbering framework has however started already at the beginning of the 90's. A several times updated Hungarian NNP as a basic technical plan for telecommunications was established and the management of numbering resources became the task of the NRA. Act XL of 2001:

- declares that numbering resources are the property of the state,
- asks for the continuous updating of the NNP (the Hungarian designation is: ANFT – Azonosítók Nemzeti Felosztási Terve, i.e. Hungarian Plan for National Identifiers),
- gives authorisation to introduce licence fees for the use of numbering resources,
- sets deadlines for the introduction of carrier selection and number portability,
- obliges service- and network operators to co-operate in the field of numbering.

The regulatory framework for numbering looks as follows:

a) Decree 75/2000 (V.31) of the Government on the management of numbering resources.

b) Decree 10/2001 (III.27) of the Minister on the National Numbering Plan (planned modifications are in progress).

c) Decree 19/2001 (X.31) of the Minister on licensing fees for the reservation and use of numbering resources.

d) Decree 250/2001 (XII.18) of the Government on carrier selection).

e) Decree of the Government on number portability (to appear in 2002).

Since 1999 the NRA pays special attention to numbering issues. The NRA took and takes part in the preparation of relevant jurisdiction. Feasibility studies by external experts have been made for the NNP, for carrier selection and for number portability. A remarkable piece of work has been devoted to establish and maintain a reliable inventory of numbering resources already in use. A part of the NRA public homepage gives information on several aspects of numbering. Last but not least the NRA is also organising the very important consultations with market players.

At present the Hungarian NNP covers telephone numbers, data DNICs, mobile IMSIs, signalling ISPCs and NSPCs. TETRA ITSIs will very likely also be included in the near future. In the case of telephone numbers the existing NNP satisfies the basic requirements of the liberalised market, but is not future proof. The present structure of NDCs, a mixture to indicate geographical areas, other networks and services has no adequate reserve. A radical change would require a rather big rearrangement with an impact on the majority of existing NSNs. This step is under consideration. In the case of short codes there is no similar problem.

8.6.6. Interworking of numbering and addressing plans

The user connected to a given network with the intention to arrive to a given service will normally use some numbering tools to access this service. From the technical point of view there are two basic situations: the service is offered either by the same or by a different network. In any case numbering tools are used to find the access point of the required service or other network in the originating network. In addition it might happen that numbering interworking supports further access and routing in the other network.

In a national PSTN/ISDN network the access point can be arrived by using the NDC part of a E.164 based number (see Section 8.6.4). If the service is offered by the same network (e.g. freephone or premium rate service) the task might be a number translation to find the routing number for the final destination (freephone subscriber or a premium rate service database). If the service is offered by an other (e.g. mobile, data or IP based) network the access point might be reached in the same way. The next step depends on the numbering plan used in the other network. Does this network use E.164 based naming (e.g mobile telephony) then only

addressing interworking is required. Does it not, interworking is necessary both on the naming and addressing level. See the example for telephone service in IP based networks in Figure 8.6.2. The considerations above refer to one phase access and can be extended to arrive at international services or networks from a PSTN/ISDN by using also the CC part of an E.164 number. If two phase access is used (e.g. an identification is required by the service or network) no further numbering interworking is necessary, since in the second step the user might already use an other numbering plan.

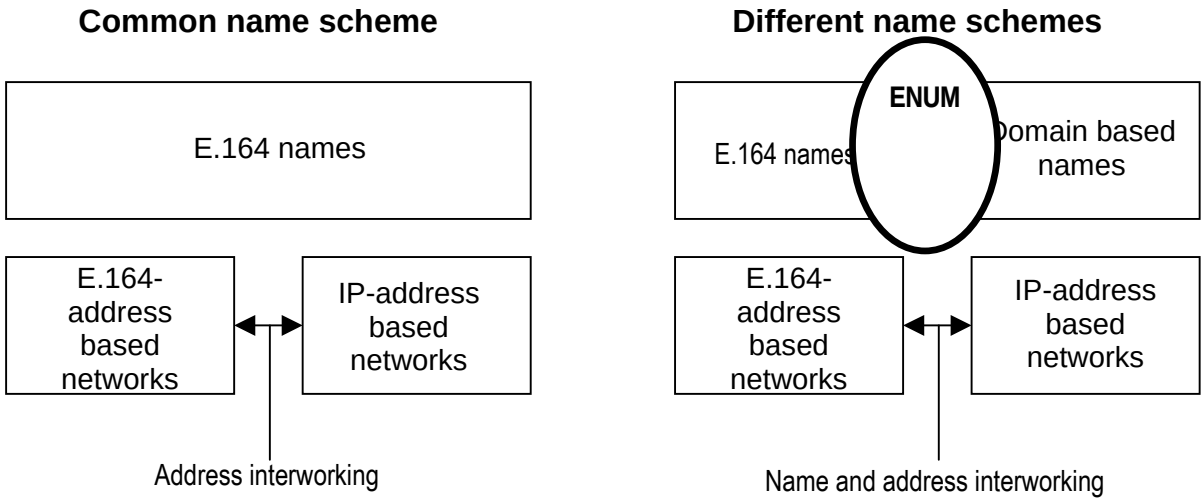


Fig. 8.6.2. Numbering interworking example

Examples for interworking of E.164 with other numbering plans can be found in Section 8.6.5. Intensive co-operation between IETF and ITU-T Study Group 2 is going on in the field of naming and addressing interworking for ENUM as indicated in Figure 8.6.2. ENUM is a mechanism for mapping E.164 numbers into Uniform Resource Identifiers (URIs) of the domain based name system corresponding to communication applications associated with those numbers. ENUM related ITU-T Recommendations, guidelines for co-operation of TSB and RIPE-NCC (*Réseaux IP Européens Network Coordination Centre*) and an explanatory type Supplement are on the way to be endorsed by the first half of 2002.

8.6.7. Number portability, carrier selection

Number portability (NP) and carrier selection (CS) are explicitly favored tools of the EU to support competition see **[8.6.8]**. The basic text of the relevant Directive

for NP asks for the introduction of number portability in the fixed telephone network between operators (a) for geographic numbers if the location of the subscriber has not been changed and (b) for non-geographic numbers in any case. The EU does not require NP if the subscriber has changed location (location portability) or service (service portability). The new set of EU directives on electronic communications having appeared in its final form in the first half of 2002 extends NP to mobile services but explicitly excludes NP between fixed and mobile networks.

In the case of CS the EU requires that at least organisations defined as having significant market power enable their subscribers, to access the switched services of any interconnected provider of publicly available telecommunications services. The subscriber might access the services either on a call-by-call basis by dialling a short prefix or by means of pre-selection. However a facility to override pre-selection on a call-by-call basis has to be catered for. In Hungary the prefix is “15cd”, where digits “cd” indicate the provider of the service to be accessed.

One has to see clearly that the numbering itself plays a minor role in implementation of NP and CS. There are many important other than numbering aspects to be solved as: handling of user requests, co-operation of network operators, decisions on billing and accounting, operation of the NP reference data base, selection of technical methods to be applied, etc. Solution of these problems seems to be the key issue.

- - - - -

There are many other important issues that have not been covered in this Subchapter e.g. the use of routing numbers, signalling aspects of numbering, the use of escape codes for numbering interworking, etc. The general picture given above, however, might help to see that numbering is a basic and not at all simple tool supporting the efficient use of telecommunication services in a liberalised environment.

References

[8.6.1] 2001. évi XL. törvény a hírközlésről. – Magyar Közlöny, 72. 2001. p. 5221-5264.

[8.6.2] Final Report on Harmonised National Conventions for Naming and Addressing – ETO, Dec. 1999, Work order Nr. 48465, Authors: J. Nuijten, M. Bernardi, pp. 107.

[8.6.3] Directive 97/33/EC of the European Parliament and of the Council of 30 June 1997 on interconnection in Telecommunications with regard to ensuring universal service and interoperability through application of the principles of Open Network Provision (ONP) – (31997L0033) Official Journal L 199, 26/07/1997 p. 0032 – 0052.

[8.6.4] Final Report on Review of Numbering Schemes on their Openness to Competition, ETO, Oct. 1997, Work order No. 48 378, Author: J. Kanervisto, pp. 67.

[8.6.5] Final Report on Harmonised National Numbering Conventions, ETO, Oct. 1997, Work order Nr. 488379, Author: J. Nuijten, pp. 54.

[8.6.6] Final Report on Harmonisation of Short Codes in Europe, ETO, Sept. 1998, Work order Nr. 48380, Author: J. Nuijten, pp. 67.

[8.6.7] Fees for Licensing Telecommunications Services and Networks – ETO, Oct. 1999, Work order Nr. 48464, Author: Ann Vandenbroucke, pp. 90 + 62.

[8.6.8] Directive 98/61/EC of the European Parliament and of the Council of 24 September 1998 amending Directive 97/33/EC with regard to operator number portability and carrier pre-selection – Official Journal L 268, 3/10/98 p.0037 – 0038

[8.6.9] The international public telecommunication numbering plan – ITU-T Recommendation E.164. May 1997, pp.25.

[8.6.10] Principles and responsibilities for the management, assignment and reclamation of E-Series international numbering resources – ITU-T Recommendation E.190. May1997, pp.12.

[8.6.11] The international identification plan for mobile terminals and mobile users – ITU-T Recommendation E.212. Nov.1998, pp.10.

[8.6.12] Assignment procedures for international signalling point codes – ITU-T Recommendation Q.708. March 1999, pp.17.

[8.6.13] Digital cellular telecommunications system (Phase 2+); Numbering, addressing and identification – ETSI EN 300 927 V5.4.1, Dec. 2000., pp.21

[8.6.14] Management of the European Telephony Numbering Space (ETNS) – ETSI EN 301 161 V1.2.1, Jan. 2002., pp.18.

8.7. Frequency Management

János Grad, author

István Hazay, reviewer

The entirety of governmental regulatory measures dealing with distribution and use of radio frequencies is called frequency management. To a more exact interpretation, it shell be defined

- what as radio frequencies qualified are,
- circumstances under them the use of radio waves enforceable are.

Radio frequency is the name of a domain of the electromagnetic spectrum with frequencies less than 3000 GHz. Not subjects of the frequency management are information transmissions based on non-electromagnetic propagations (e.g. underwater ultrasonic communication) further on electromagnetic communication systems with frequencies higher than 3000 GHz (infrared, light, ultra-violet and X-ray transmissions).

A regulatory framework by frequency management is given to radio waves propagating in the free space only. There are no frequency management criteria to procedures applied inside equipment and to wireline signal transmission. At the same time, spurious emissions of equipment and telecommunication lines are regulated versatilely. It can be said that the subject of frequency management is the use of free propagating radio waves.

8.7.1. Interferences and their prevention

Why is there a need for regulations in the case of free propagating radio waves? The reason of this is the possible appearance of radiotechnical disturbances, interferences for lack of regulations or in case of infringement of the existing regulations. Interference is a disturbance occurrence where

Interference is a phenomenon of disturbance where signals of other radio sources get in any receiver point of a radiocommunications link as the proper signals and the false signals take effect on the reception.

In theoretical respect, all kinds of interference would be avoidable if each frequency was used only once in all over the word. In this case, it would be attainable that beside the usable signal even the smallest alien signals couldn't get in the input of a receiver. However, signals of distant sources are not disturbing in practice. It is raised how to determine when a signal disturbing is. Therefore the criteria of disturbance (interference) shall be defined.

In radio reception, the interference relations depend on the properties of the radio connection and the disturbing signals, in particular on the magnitude of the signals. The transmission properties (e.g. modulation, coding, access method), parameters of receiver (sensitivity, receiver mask) and form-characteristics of disturbing signals (impulse filling-in factor) have a share in evaluation of disturbance.

Interference can be characterized by well-defined parameters. Obvious that not all kinds of alien signal influence produce harmful interference. The international practice defines standard values for interference criteria that are thresholds separating acceptable interference and harmful interference.

8.7.2. Radio services and their allocations in the radio spectrum

The International Telecommunication Union (ITU) has grouped the radio applications in big categories, in radio services, which differ from each other characteristically in the type of application and in interference properties. Some example for radio services:

- Broadcasting service – conveying of a radio programme from a terrestrial transmitter point to numerous receivers;
- Broadcasting satellite service – conveying of radio programme from a transmitter point onboard satellite to terrestrial receivers;
- Fixed service – terrestrial radio communications systems with fixed site stations;
- Fixed satellite service – radio communications systems where fixed site earth stations are connected to each other with satellite.

The terrestrial radio services often have satellite equivalent. In general, the above radio services can be put in pairs to each other but there are several satellite

services we can’t arrange similar pairing. For example, there is no pairing at the following satellite service:

Intersatellite service;

Earth exploration satellite service.

Sometimes there is no use to distinct radio services using terrestrial and satellite techniques. An example to this is the radio astronomy.

To a simpler handling of interferences, the ITU arranges the radio services in blocks (bands). In this way, a characteristic frequency allocation table comes into being. One or more radio services are associated with the individual frequency bands. For example:

8400 – 8500 MHz Fixed service Mobile service (expect aeronautical mobile) Space research (space-to-Earth)

8500 – 8550 MHz Radiolocation

8550 – 8650 MHz Earth exploration-satellite service (active) Radiolocation Space research (active)

etc.

If more radio services are located inside a frequency band, these are chosen in such a way that their effect to each other is to be well evaluated and there is a theoretical possibility to creating criteria for interference protection. Beyond technical criteria, there are administrative means to the reduction of interferences between radio services. An administrative is the priority associated to the services.

A radio service can have primary, secondary or tertiary priority. If there are both primary and secondary services in a frequency band so a secondary service can’t cause any harmful interference to the services with primary priority and they can’t claim any protection from interferences of a primary service.

Tertiary services are called non-interference services. A service like this can’t claim protection again interference at all and it can’t cause any interference to other services with higher priority.

It is a frequent case that more services have the same priority (mostly primary priority) in the same frequency band. In such a case it can’t be decided generally

which one has an advantage over the others. The preference for use is ordered to the given realized radio systems. In case of services with the same priority, that radio application (system) of radio service has preference over others which was deployed earlier, more exactly, whose registration has an earlier date. A preference belongs to the user registered its claim earlier (compliance of demands in order).

The radio band stretch up to 3000 GHz but the part of 9 kHz – 275 GHz is allocated among different services. Radio frequencies not falling into the above range, the very small and very high frequencies from the point of view of radio communications, for the time being can be used without restriction for radio applications in any radio services.

The radio allocation is global in many of frequency bands so they can be used worldwide. However there are numerous frequency bands having regional allocations. Considering the radio applications, the world is divided into three regions:

- Region 1: Europe, Russia, Africa and Near-East,
- Region 2: America
- Region 3: Middle- and Far-East, Australia and Oceania.

In the regional allocation of radio services, the different demands and requirements of the single regions are reflected.

8.7.3. Word Radio Conference, Radio Regulations

The global and regional allocation of the radio services in the radio spectrum is based on a compromise agreement of the countries of the world. The venue of negotiations is the series of Word Radio Conference organised in 3 years. The last one was Istanbul, Turkey, in the year 2000. The next Conference will be arranged in Caracas, Venezuela, in 2003. The Word Radio Conferences are organised by the ITU. All the countries have right to participation and almost all of them take part in it.

The International Telecommunication Union (ITU) organizing the World Conferences is the oldest professional organization of the UNO. It was established in 1865 and Hungary is among the 20 establishing country.

The summary of decisions and resolutions of World Radio Conferences is the Radio Regulations (RR). Each World Conference accomplishes the RR with renewed

amendments. So the RR became a collection of agreements had came into being through many ten years.

By size, the RR is a many-volume document. Its most important part is the frequency allocation table, disposing of allocation of radio services in the radio spectrum. In further parts, it deals in detail with the power restriction of radio services, as this is a provision of co-existence of different services.

The RR is a treaty-document based on mutual agreement of numerous sovereign states. The agreements have been concluded by a chain of compromises. Special means of compromise making are the *footnotes*. These afford possibility to individual countries or groups of countries to deviate from the general solutions (e.g. there are footnotes applying by the EU-countries for the provision of their special telecommunications needs). If a country is not concerned in a footnote, it is very difficult to join it later because a precondition of joining is the acceptance of all neighbour countries. (Hungary has seven neighbours; it needs 7 statements of acceptance.)

The RR has been signed by delegates with governmental authorization and it becomes part of the national law and order of the countries. That's so in Hungary. RR is a substantial guide of the Hungarian radiocommunications.

8.7.4. Recommendations, standards and agreements in radiocommunications

Recommendations

Two big international regulatory systems are in use in Europe based on the RR. One of them is a system of documents has been worked out by study groups of the ITU. The other is a collection of documents has been created by the regional telecommunications organisation of the European countries, that is CEPT. Concerning regulatory documents are called '*Recommendations*' in general because this is not an interstate treaty based on consensus, but regulatory criteria worked out in expert's committees. Acceptance of the Recommendations is facultative but strongly recommended, as hereby a harmonised radio use has been reached.

Considering their regulatory effects, the Recommendations can be divided into two groups.

Fact-describing Recommendations belong to one group (e.g. what type of modulations is recommended to use and what parameters of different modulations should be taken into account, etc.) Fact-describing recommendations are very useful from technical aspect. Recommendations issued by ITU include the entirety of the radiocommunications and they give together a complete monograph of the whole radiocommunications.

Regulatory Recommendations represent the other big category (e.g. the collection of rules describing the creation of frequency-channels in the radio transmission). Regulatory Recommendations are based on the Fact-describing Recommendations, which give the technical background. That's why the fact-describing and regulatory Recommendations together provide a common background to the radiocommunications.

By their structure, ITU-Recommendations and CEPT-Recommendations are identical but there is a characteristic difference between their focuses. At the ITU, the number of the technical type fact-describing Recommendations is higher, whereas at CEPT-Recommendations, the regulatory Recommendations are of overwhelming majority. Due to this, the CEPT-Recommendations are of big importance to the European countries from the point of frequency management but they don't have too big influence on the technical methods. However, CEPT-Recommendations have a fundamental additional function. These are means to the European Union to taking effect on CEPT-countries outside the EU, regarding the radiocommunications.

A characteristic feature of CEPT-Recommendation is that they give guiding to licensing and operation of the European-type radiocommunications systems like the GSM radiotelephony system, the UMTS 3rd generation mobile telecommunication system, the TETRA dispatcher system, HIPERLAN local radio network and the European satellite systems for telecommunication and navigation.

Standards

A further tool of the regulations is the ***standardisation***. The European organisation of it is the European Telecommunication Standard Institute (ETSI). The

European standards are published by name ETS (European Telecommunications Standard) and EN (European Norm). EN-standards of the telecommunications, together with other EN-standards referring to other areas) represent parts of the standard-system of the European Union. Standards issued by ETSI specify the operation of telecommunication systems and the minimum requirements for the characteristic parameters. It shall be emphasized ETSI standardizes not equipment but systems. However, the given standard-prescriptions often affect by the applied equipment, so they result in equipment standards as a secondary product of the system specification.

The work of ETSI can be associated inter alia with working out the standards to the GSM and UMTS mobile systems as well as specifying most types of European telecommunications systems.

Standard are mandatory neither in the EU nor in Hungary. However, there are some essential radio parameters can't be facultative because this would result in harmful interferences. That's why the CEPT in its Decisions imposes the use of specific standards or the substantial parameters of these standards. National implementation of the Decisions is not mandatory; nevertheless a country declares its join. For countries having implemented Decisions (e.g. all of the EU-s countries), parameters defined in the regarded ETSI-standards are obligatory.

Agreements

There are specific radio applications where the usable frequencies or frequency blocks are defined in international treaties. An example to this is the Stockholm Agreement on the definition of broadcasting frequencies.

In other cases, international treaties prescribe what is the condition of coordination of radio use between neighbour countries. A treaty for coordination is the Vienna Agreement that defines the conditions of coordination for applications in the fixed and mobile service. Beyond the above and other similar multilateral agreements, bilateral treaties are of high importance to Hungary that has seven neighbours. Interference-less near-border radio operations between Hungary and its neighbours are supported with bilateral agreements. However, these agreements mean significant restrictions for radio operations in the area of borders.

A question is raised, what '*area of border*' means. The distance of coordination i.e. the distance from country border where coordination is needed depends on the specific radio application and the frequency band. Its actual value is the objective of countries. For small countries, the coordination area can cover a big part of the country. In case of numerous radio applications in Hungary, there is no point in the country, where coordination with a neighbour country wouldn't be needed.

8.7.5. Legal regulations in Hungary

For the time being, the legal basis regulating the common rules of frequency managements and operation of regulatory boards is given in an independent act (Number LXII. Act on the Frequency Management in year 1993). From the nearest future, the Telecommunications Law will be enacted and it will provide the basic regulations.

Frameworks of the Hungarian radio regulations are defined considerably by international agreements. These frameworks shall be filled in according to national particularities. A substantial document of the Hungarian radio communications is the National Table of Frequency Allocations (NTFA), which was issued, in form of governmental decree so it is a part of the Hungarian law and order. Structure of NTFA corresponds to structure of Radio Regulations (RR) It contains the Hungarian frequency allocations of radio services, divided into frequency bands. Beyond this, NTFA creates another breaking-down on what radio services in different frequency bands in civil either in governmental or in common use are. Governmental use is called in Hungary the radio use the armed bodies.

In the NTFA beside the radio service indication of frequency bands, national footnote can often be found. National footnotes specify the concrete use of the given band in Hungary. Footnotes declare what radio applications can be realised inside a radio service in Hungary. In the footnotes, there are often cited ITU and CEPT Recommendations and further technical prescription related to the relevant applications.

It is always shown in footnotes whether the state of a radio application is assignable, reserved or planned. Licence can be granted to a radio application being in assignable state. Reserved radio applications may get to assignable state in a

short time in case of proper conditions. A declaration of this doesn't need any rule of law; it is enough a statement from the competent supervisory authority. The planned state provides instruction on the larger timescale direction. Change from planned state to reserved or assignable state needs a modification in the NTFA so it is a rather complicated process.

NTFA as a governmental decree is the highest-level legal regulation of the entirety of radio bands in Hungary. A detailed legal regulation of the respective bands is contained in ministerial decrees. The NTFA and ministerial decrees form legal means allowing inclusion of international treaties into the Hungarian right.

Three categories shall be distinguished precisely during the use or granting the subbands of the frequency spectrum. These are the *allocation*, *allotment* and *assignment*.

Allocation: *Entry of radio services in the frequency table of the Radio Regulations or NTFA.* Generally, the radio services in the allocations cover numerous different radio applications.

Allotment: Definition of the forms and channelling plans of licencable radioapplications in a frequency band allocated to one or more radio services.

Assignment: *Licencing an application in conformity with the allotment*. The licence defines the operating frequencies or channels and other operational preconditions.

8.7.6. Regulation not based on law

The Communication Authority may sometimes create regulatory documents that are not law. These are the frequency utilisation rules used by the Authority in the process of frequency assignment. In many cases, international agreements don't get legal background but they will be issued as Authority recommendations.

The regulations on lower level then law and issued by Authorities always have a well-founded background regarding its standardisation, compatibility and international relations. The licensing authority acts on these bases issuing administrative decisions. These documents are mandatory to the licensing authority and but informing only to the clients.

The experience shows that these lower level regulatory documents are very useful tools at regulating the process of the frequency assignments. These enable an effective and quick licensing procedure.

If a client appeals against an administrative decision based on information documents, the Authority can review the matter on the basis of the valid law only. This often needs domestic interference evaluation or carrying out the complete process of international coordination. A procedure like this takes a lot of time and one can reckon with a refusal from abroad. A legal remedy for an unwanted result of international coordination is already an objective of international rights.

The international coordination forms one of the basic task of the national radio administrations. Tasks to perform in this scope of duty are

- notification of Hungarian radio stations for acceptance to the radio administration of the respective country,
- evaluation and approval or refusal of the requests arriving from abroad.
- notification of stations, requiring international protection, to the ITU Radio Buro where they will be recorded into the *Master Register*.

8.7.7. Methods of frequency assignment

In general, frequency is qualified as restricted natural resource. A limited bandwidth is available to the single radio applications and satisfaction of demands shall be solved not to produce harmful interference between existing and further radio systems.

Degree of limitation of the frequency resource is different by radio services or applications. Frequency set of area-covering radio supplies (e.g. broadcasting, cellular mobile telephony, etc.) can be used only once at the assigned territorial item and can't be repeated in the same area. The fact that the restricted frequency resource is scanty for area-covering applications is unambiguous. In contradiction to this, high directivity antenna characteristics reduce the risk of interference in such a degree that the same set of frequencies will be applicable to a numerous connections in the same area. Thus a large number of demands can be satisfied despite the set of frequencies is limited so the resource can't be taken as scanty.

The method for the assignment of the elements of the frequency set is, if there are no other legal measures, fulfilment of the applications in order. In this case the

frequencies of the set shell be assigned till they are available. After that, further assignment is not possible.

In case of scanty restriction, when small number of elements belongs to the frequency set, the fulfilment of demands in order can hardly be taken into account. Instead of this, the applicants shall compete somehow for the required frequencies. There are several types of the competition for assignments. Typical procedures:

- Tender – recently this is the method for assignment in broadcasting in Hungary.
- Concession tender – in the preceding years this was the tender procedure for mobile telephone services.
- Frequency auction – price bidding after unambiguous declaration of technical and service parameters. So the only parameter of the competition is the price.
- 'Beauty contest' – a competition that takes into account a lot of parameters of service and technical features at the same time. Weighting among parameters is a subjective thing; this is referred by the name 'beauty contest'. The law makes possible this type of assignment in Hungary from the year 2002.

8.7.8. Licences authorising the use of frequency

There are two fundamental types of licences in connection with each radio licencing procedure: Frequency assignment decision and radio licence.

The frequency assignment decision means substantially a licence to deployment. In this decision, the authority assents to set up radio stations in geographical sites or areas included in it. A frequency assignment decision is based on preceding planning, on the so-called frequency plan.

A frequency plan is a telecommunications-engineering document defining the characteristics of transmission and the radio parameters of the planned connection. The plan shall describe inter alia the relations of interference in two respects. On the one hand, it shall be studied what interference effects the new network takes to systems having been earlier deployed or licensed. On the other hand, it shell be studied how the planned network bears interference from the existing radio surroundings. Thus, the frequency plan contains all of essential parameters of the given network from the point of frequency management.

A frequency assignment decision based on frequency plan gives right to realisation and switching on the radio network further on to carry out measurements before the regular operation.

Radio licence concerns the operation of radio networks. Its technical content it comprises the real (measured, adjusted) parameters of the realised radio systems.

A special form of radio licences is the so-called ’general licence’ that can be issued if the licensed parameters of the related radio equipment can’t be bound to individual structure or defined sites of the radio networks. Examples for general licences are the terminals of mobile communications and numerous types of low power equipment (alarm equipment, cordless cameras, etc.).

Another special type of licences is the experimental licence. This licence, suitably to its name, may be granted for a strongly limited time period in order to carrying out well-defined experiments.

The authority provides data delivery in form of official decision. It seems to be useful a special kind of radio licences, the principled building licence. Here, together with data delivery, there are given additional conditions of the frequency plan to be realised. It provides a security that the Authority will accept the plan, presuming that the planning is correct.

It is important that the operation of radio sources (stations) shall comply with the requirements stated in the licences. Any deviation from the terms in the licence during the operation or any possible stations working without licence may jeopardize other stations by interference. An important task of the authority is the continuous inspection of the radio activity in the whole territory of the country. The inspection has administrative and technical means. Among the technical means, the *spectrum monitoring* is of fundamental importance. It means the observation of the frequency, the power of the radio transmissions, the distribution and occurrence of them.

8.7.9. Frequency fees

Frequency fees shall have two fundamental functions:

- They shall render an indicator on the utilisation of the restricted frequency resource among several circumstances.

- They shall provide revenue, proportional with utilisation, to the state for covering expenses of tasks of authorities', relate to communications.

Kinds of frequency fee:

- Frequency occupancy fee, due one-time – determined by frequency assignment decision.
- Frequency utilisation fee – its amount is determined in radio licences.

In general, the law disposes the method of establishment of frequency fees expect of cases where the service is based on concession contract. In this case, the frequency fee is determined in the concession contract.

The theoretical basis of determination of frequency fees is the occupied bandwidth by radio channel and, indirectly, the size of areas where the frequency band is not usable by others because of the given radio application. The fee depends even on the type of radio applications and on the location of the actual frequency band inside the radio spectrum. In general, lower frequencies are more valuable.

By introduction of frequency fees, the regulatory system of frequency management, based on technical principles and built on authority procedures, is supplemented with economic type, normative regulation.

References

[8.7.1] Radio Regulations, Edition 2001. Geneva 2001.

[8.7.2] Government Decree No. 221/1999 (XII. 29.)
On the Establishment of the National Table of Frequency Allocations

[8.7.3] Act LXII of 1993 on Frequency Management

[8.7.4] Decree No. 6/1997 (IV.22.) KHVM on frequency occupancy and utilisation fee

[8.7.5] ITU Recommendations: www.itu.org

[8.7.6] Green Paper on radio spectrum policy
European Comission, 1998.

[8.7.7].CEPT Recommendations and Decisions: www.ero.dk

[8.7.8] ERC Report 25: Frequency Range 29.7 MHz to 105 GHz and Associated European Table of Frequency Allocations and Utilisations

[8.7.9] ETSI Standards: www.etsi.org

8.8. Application security – electronic signature

István Rényi, PhD, author

Iván Schmideg, PhD, reviewer

Information society will be built upon open public communication networks. One of the key elements of the establishment of a streamlined public administration, and the competitiveness of our national economy will be the effective usage of on-line transactions in the years ahead. Electronic documents, having the same legal validity as conventional ones, electronic contracts, electronic bank transfers, and e-commerce are rapidly gaining importance. Message transfers, which often take place between parties previously unknown to each other, could not be carried out without secure technologies and suitable legal regulation. An authentic transfer is one, in which the receiving party is sure of the authenticity of the signer, the integrity of the message, and that the signer cannot later repudiate having sent the message. All three requirements can be met simultaneously by using electronic signatures, based on *Public Key Infrastructure (PKI)* technology [8.8.1]. Message confidentiality is not handled by electronic signatures, therefore, encryption is not part of our discussion.

8.8.1. Technical background of electronic signatures – PKI

To understand the technical process involved, two procedures have to be made clear first: The first one is *encryption*, which yields a scrambled message from a plain text message using a cryptographic algorithm, governed by a key value, which is unique. The basis of all algorithms used here is a so-called mathematical hard problem, which is easy to compute in one direction, but which is very difficult to invert [8.8.2, 8.8.3]. The most widespread version of reliable, standardised signature algorithms is RSA (named after the author trio Rivest-Shamir-Aldeman) [8.8.4]. The basic mathematical problem here is the following: it is easy to multiply two large prime numbers, but it is difficult to factorise the product. Another algorithm, DSA (Digital Signature Algorithm) [8.8.5] is also used extensively. DSA's mathematical hard problem is that it is easy to raise to a power in the multiplicative subgroup of a finite body, but to produce the inverse, the discrete logarithm is hard. The use of

ECDSA (Elliptic Curve analogue of DSA) [8.8.6] is rapidly spreading, because of its far less computational requirement. The basic math problem: for the points of elliptic curves fulfilling some special requirements, addition can be defined, for which the points of the curve make up a group. Multiple application of this operation (scalar multiplication) is easy, whereas the inverse operation, performing discrete logarithm is difficult.

The second procedure to note is making a fingerprint (a compressed value) of the message, by a hash function. Compression is necessary because the secure signature algorithms used can only sign short electronic messages in a reasonable amount of time. A hash function is one, which transforms an input bit-stream of arbitrary lengths into an output stream of fixed lengths, having the following two properties:

- it is practically impossible to find an input which would be transformed by the hash function into a given output (preimage resistance),
- it is practically impossible to find two distinct inputs which would be transformed into the same output (collision resistance).

Regarding hash functions, two algorithms are accepted for use in electronic signature standards: **SHA-1** (**S**ecure **H**ash **A**lgorithm) and **RIPEMD-160** (**R**ace **I**ntegrity **P**rimitives **E**valuation **M**essage **D**igest Algorithm) [8.8.7].

Public Key Cryptography (PKI), the most widely used electronic signature technology today, employs two different but mathematically related keys (an asymmetric key pair): a private key (usually called: signature-creation data) for creating the signature, and a public key (usually called: signature verification data) to verify it. A message encrypted by one of the keys can only be decrypted by the other one. An important property of the key pair is that it is computationally unfeasible to derive one key from knowledge of the other one. Furthermore, the encryption algorithms are such that the encryption key cannot be reproduced from knowing both the plain and the encrypted messages.

The signing process includes the generation of the hash value of the original message, which is then signed by the signer’s private key. This is called *digital signature*, which is appended to the original message.

During verification the following steps are carried out:

1. the hash value of the original message (h) is generated with the same hash function that was used by the signer;

2. the digital signature is decrypted with the signer's public key to recreate the hash of the original message (h');

3. the outputs of step1 and step 2 are compared.

If the two outputs match (h = h'), it can be reasonably assumed, that the contents of the message have not changed during transmission, and also that the signature was made with a private key, the counterpart of which is the public key used in step 2.

8.8.2. Certificates and Certification-Service Providers

To verify the authenticity of a signature the relying party needs to use the public key of the signer, and also should know who is in possession of the counterpart private key. In PKI technology a *certificate* is used for this dual purpose. The certificate binds the identity of the signer in possession of the private key with the public key. The certificate is issued and signed, in an unforgeable manner, by the *Certification-Service Provider (CSP).* The service provider has its own certificate, which is usually signed by another service provider. This way, a chain or a hierarchy of trust is established. Certificates can be retrieved and their validity checked by accessing to the service provider's directory service, and – to speed up the verification process – it is usually attached to the body of the message.

The data structure of the X.509 type certificate [8.8.8] used exclusively in PKI includes the name (or alias) of the signatory, his/her public key, details of the hash- and signature algorithms, the validity period of the certificate, the identity, the public key and the employed algorithms of the Certification-Service Provider.

The relying party verifies the authenticity of the certificate (by examining the electronic signature of the service provider). If it is found authentic and originating from a service provider he trusts, then it can be reasonably assumed that the private half of the key pair (the public half of which is stored in the certificate) belongs to the individual in possession of the certificate.

Identification of the certificate user/owner is carried out by *Registration Authorities* that operate under full responsibility of the Certification-Service Provider.

The procedures and the list of required documents may differ for certificates of different security levels.

If a certificate for any reason becomes invalid, the service provider has to revoke it in a timely manner – based on an authorized and validated revocation request. Such a reason could be, e.g. the expiration of validity of the certificate, if the certificate contains invalid data, or if the subscriber's private key has been lost, stolen or potentially compromised. For this reason, the CSP shall maintain a *Certificate Revocation List (CRL),* and publish it at least daily over open networks. Upon verifying a signature, the relying party should check whether the certificate was valid at the time the signature was created.

Often, an authenticated time should be bound to an electronic transfer. This may be necessary, for instance, to prove the authenticity of an electronic document, to decide whether the signature was placed before or after the certificate was revoked. For this purpose a *time stamp* is used, which is issued by another trusted third party, the *Time Stamping Authority*. Upon receiving the hash code of the document from the signatory, the Time Stamping Authority attaches a time stamp token to it, generates a new hash code, signs it with its own signing key, and sends the whole thing back to the signatory.

The signatory is required to handle his private signing key in such a way that no one else could have access to it, and sign with it in his name. Keeping the private key on a flexible or a hard disk is not secure, this method can only be used in a safe environment, e.g. at home. Reasonable security is achieved by using a smart card, activated by a PIN code or some biometric method (fingerprint, iris image, voice, etc.) to store the key on. The microchip on this device has its own processor, memory, and operating system, it stores the private key in a highly protected manner (it can never be accessed from outside), and it may also store the user's certificate. The smart card runs the cryptographic signing algorithm, i.e. on its output it provides the digital signature of the document hash code, fed to its input. Newer cards may store multiple keys (and run multiple applications), some include hash code generators, random number generators, and are capable of on-board key generation. The latter capability excludes the possibility of anyone making a copy of the private key and impersonating the card owner.

The key pair is either generated by the subscriber (on the smart card or by a user program, like a browser) or, alternatively, a service provider may prepare and provide the *Signature-Creation Device (SCD).* Examples of the Subscriber-Device Provision Service are:

- A service which generates the subscriber's key pair and distributes the private key to the subscriber;
- A service which prepares the subscriber's Secure Signature-Creation Device (SSCD) and device enabling codes and distributes the SSCD to the registered subscriber.

8.8.3. Legal background

The legal status of electronic signatures should be determined by each country, in its own legal system. However, because of the global, cross border usage, it is necessary to base the legal recognition on the same principles, to have the same guarantees for security, and to make sure the systems are capable of working together. For this reason, the European Parliament and the Council issued a Directive (1999/93/EC) [8.8.9] on a Community framework for electronic signatures in December 1999. The Directive was to be implemented in 18 months by the Member States. Because of the harmonisation requirements of its legal system, Hungary also adopted the Directive. The Electronic Signature Act (XXXV/2001.) [8.8.10], which entered into force along with three lower level regulations on September 1, 2001 is fully compliant.

The most important element of the regulation is legal recognition, i.e. legal effectiveness of an electronic signature and its admissibility as evidence in legal proceedings cannot be denied only on the grounds that it is in electronic form. This general acceptance criterion is true for any kind of electronic signature, even for those without any security constraints. The Electronic Signature Act defines the *advanced electronic signature*, which meets the following requirements: (1) it is uniquely linked to the signatory and is capable of identifying him; (2) the Signature-Creation Device is under the sole control of the signatory; (3) it is linked to the data in such a way that any subsequent change of the data is detectable. If "written form" is required by any regulation, this can be satisfied by a document signed with an advanced electronic signature. The highest quality signature is the *qualified electronic signature, which*, in addition to being an advanced signature, is based on a

qualified certificate and is created by a Secure Signature-Creation Device (SSCD). This kind of signature attached to a document satisfies the requirements of a handwritten signature.

According to the law, electronic signature services are the following: certification-services, time-stamping services and subscriber-device provision services. These services may be provided either one-by-one or combined.

According to the Directive Member States may introduce *voluntary accreditation schemes* aiming at enhanced levels of certification-service provision. All conditions related to such schemes must be objective, transparent, proportionate and non-discriminatory. Each Member State shall ensure the establishment of a system that allows for *supervision* of Certification-Service Providers, which are established on its territory and issue qualified certificates to the public. The voluntary scheme has not been implemented in Hungary. The Communication Authority Hungary, a governmental organisation, is designated by the law for registering all service providers and for the mandatory supervision of qualified service providers. Besides these, the Communication Authority shall register all electronic signature devices and other products, which can be used for the generation and verification of advanced and qualified electronic signatures.

Requirements for qualified certificates, requirements for CSPs issuing qualified certificates, requirements for SSCDs and recommendations for secure signature verification are given in the Annexes of both the Directive and the Hungarian electronic signature act. Based on these requirements, a number of standards are being worked out. The activities, supported by the European Commission and carried out by the European Electronic Signature Standardization Initiative (EESSI) [8.8.11], are focusing on interoperability issues and harmonised security requirements.

8.8.4. Security issues

The major weakness of electronic signatures and the main cause of most legal problems is the clear physical separation of the signer and the signing device. To avoid unauthorized signing, the owner of the private key must protect the signing device (and its activating PIN code) by all possible means from getting into the

possession of another individual. Should this happen, however, the certificate has to be revoked without delay.

Both the signer and the verifier are required to perform the signature creation and verification in an overall secure environment. Otherwise, it cannot be guaranteed, that the signed data is equivalent to the one displayed on the screen, or that the verifier won't get a false "valid signature" response. The EESSI documents mentioned above provide a good starting point to implement/deploy a secure signature-creation and -verification environment.

To maintain a secure, abuse free operation of the electronic signature system, to create and maintain the trusted environment, the trustworthy and reliable operation of the CSP is of primary importance. Supervision of service providers issuing qualified certificates is carried out by the Communication Authority Hungary. The assessment is based upon the electronic signature law and its lower level regulations, and those technical specifications and requirements that follow from the legal regulations. The security requirements, against which the compliance is assessed, are classified into two categories: the first category includes procedures, internal operation, management and processes of the service provider, laid dawn in the *Certificate Policy (CP)*. The second category includes baseline environmental (physical) and IT security requirements. Supervision is carried out by the Authority on a yearly basis, unless exceptional situations (like a complaint) require attention. Certification-Service Providers are mandated to publish their *Certification-Practice Statement (CPS)* and Contractual Terms and Conditions for their clients by a durable means of communication in readily understandable language.

Abbreviations

PKI	–	Public Key Infrastructure
RSA	–	crypto algorithm, worked out by Rivest, Shamir, Aldeman
DSA	–	Digital Signature Algorithm
ECDSA		Elliptic Curve DSA
SHA	–	Secure Hash Algorithm
RIPEMD		Race Integrity Primitives Evaluation Message Digest Algorithm

CSP – Certification-Service Provider

CRL – Certificate Revocation List

CP – Certificate Policy

CPS – Certification-Practice Statement

References

[8.8.1] Public Key Infrastructure (X.509) (PKIX),

[8.8.2] B. Schneier, Applied Cryptography, Wiley, 1997.

[8.8.3] "Regulierungsbehörde für Post und Telekommunikation, Geeignete Kryptoalgorithmen, 5.7.2001, at http://www.regtp.de/imperia/md/content/tech_reg_t/digisign/29.pdf [8.8.6] Johnson, D.B., and Menezes, A.J., "Elliptic Curve DSA (ECDSA): An Enhanced DSA, " at http://www.certicom.com/research/wecdsa.html

[8.8.4] Rivest, R., Shamir, A. and Adleman, L., "A method for obtaining digital signatures and public key cryptosystems," *Communications of the ACM*, Vol. 21, No. 2, pp. 120-126, 1978.

[8.8.5] S. Burnett, S. Paine, RSA's Official Guide to Cryptography, McGraw-Hill, 2001, pp. 160-163

[8.8.6] Johnson, D.B., and Menezes, A.J., "Elliptic Curve DSA (ECDSA): An Enhanced DSA, " at http://www.certicom.com/research/wecdsa.html

[8.8.7] Dobbertin, H., Bosselaers, A. and Preneel, B., "RIPEMD-160: A strengthened version of RIPEMD," in Fast software encryption, Proceedings of the Third International Workshop, Cambridge, UK, February 21-23, 1996

[8.8.8] Public Key Infrastrucrure (X.509) (PKIX), http://www.ietf.org/html.charters/pkix-charter.html

[8.8.9] European Directive on Electronic Signature 1999/93EC, http://www.ict.etsi.org/eessi/e-sign-directive.pdf

[8.8.10] 2001. évi XXXV. törvény az elektronikus aláírásról, http://www.hif.hu/es/laws/eatszoveg.pdf

[8.8.11] European Electronic Signature Standardization Initiative http://www.ict.etsi.org/eessi/EESSI-homepage.htm

9. What comes next?

Miklós Boda, author

György Lajtha, dr., reviewer

Mobile Internet, as the communication system of the future has become a frequently used expression these days. During the past decade, mobility and the general spreading of mobile telephones were the biggest success in telecommunications, while datacommunication became part of our everyday lives through the spread of the Internet. However we can notice the convergence of telecommunications and datacommunications. It is hard to answer the question “How do we want to and how will we communicate in the future?”, since all that we know for sure about future is that it will be different from what we imagine at the moment. However we expect that the past decade’s two massive hits, mobility and Internet will play a part also in future’s telecommunication systems. The first steps have already been taken: during the past few years telecommunications and datacommunications have been converging. Now you can make a telephone call over the Internet, although there is still a lot of work to be done until it meets the high standards characteristic of telecommunications. When designing the next generation of mobile telecommunication systems the main objective is to enable data transmission in a mobile environment. This may hint at the fact that Mobile Internet will require a large bandwidth and a quality higher than that of today’s best effort services.

Today, mobile communication is primarily signified by GSM technology, but there are certain countries, such as the USA and Japan where different mobile systems, based on other standards, are general. This limits mobility, as GSM mobile telephones can only be used in countries that use GSM standards. Therefore it is very important for the future communication systems that the whole world uses the same standards and principles. Third Generation Mobile systems (3G) are designed to offer several features of the Mobile Internet. The

standardisation of 3G systems is performed by 3GPP, and is aimed at being accepted globally, after all this could be the key to the realization of the Mobile Internet.

We have already seen similar attempts aimed at the unification of the world's telecommunications. In the early 90s telecommunication researchers and developers were excited about the transfer to B-ISDN (Broadband Integrated Services Digital Networks) technology. According to the plans B-ISDN would have replaced the telephone network, at the same time promising the possibility of several new applications beside voice transmission. The concept included transmission of high-resolution still and motion pictures, videophone and data transmission. Based on the standards still valid, B-ISDN networks would have been based on the ATM (Asynchronous Transfer Mode) network protocol. Under the ATM layer the standard pictured a logical network built on optical cables with mainly SDH (Synchronous Digital Hierarchy) technology.

By now we can see that the world has passed by the original ideas of B-ISDN, without realizing them. The reason for this in all certainty is not the lack of demand, since during the one and a half decades elapsed since the design of B-ISDN, a global wide-band transmission network came to life, offering almost exactly the same services that had been planned to be the tasks of B-ISDN. By now apart from electronic messaging and data-base access, Internet can be used for making phone calls, what is more, various multimedia applications keep appearing all the time. While B-ISDN's planned Video On Demand service (a video library over the network) stayed on paper, the music-exchanging programs offered by the Internet gained ground in no time and gave the big music publishers plenty to think about.

The failure to realize the B-ISDN conceptions cannot be explained by the unsuitability of the selected technology either. Since today ATM is one of the important elements of Internet, replacing previous solutions in more and more networks. This phenomenon, however, properly indicates the basic difference between B-ISDN and Internet. When designing B-ISDN the objective was to

describe the entirety of a telecommunication network, from physical layer, through network protocol to applications. The first logical step of this design method was to determine future services that needed to be taken into consideration in every further step of the design process. The elements of the standard were fitted together so that they served the ultimate purpose – the application provided for the user – in the best way possible. Advantages of this vertical approach are evident, supposing that we really know future applications in advance and we can choose all elements of the system we are about to build. As a result of the linking of design steps, however efficient is the system in the conditions originally planned, it is just as inflexible when certain elements have to be modified because of the changing of the basic conditions (e.g. emergence of new services).

Internet is structured in accordance with horizontal design principles, radically differing from the above method. Each protocol level is designed to provide simple services for the layer above, thus ensuring independence of the application. This explains why Internet hardly achieves the service of quality of telecommunication networks, but at the same time surpasses by far the traditional telecommunication networks as regards flexibility.

Horizontal design principles result in the basic design attribute of the TCP/IP protocol family, its so-called end-to-end nature. The end-to-end principle means that a significant amount of network functions are delegated to the terminals. For example in end-to-end networks encryption, fault-free transmission or reliable delivery of messages constitute the responsibilities of the terminals, not those of the networks. In such networks the network itself performs dull and simple data transmission, while terminals are intelligent, and it is their task to perform difficult operations.

This principle is diametrically opposed to the design principles generally used in telephony or conceived in the case of B-ISDN. There, terminals are simple and it is the network that provides the difficult services. Call transfer, short numbers or automatic redial when a busy line is disengaged are such services.

This sort of design is extremely advantageous for the network operator from a business point of view, since the network operator decides about the introduction and the charging of the services. In an end-to-end network the above functions are available from the terminal, therefore they cannot bring extra profit.

In a certain sense every network that provides data transmission services is automatically of the end-to-end character. In this way the spreading of fax machines means the end-to-end utilisation of the telephone network, where the added value of communication is provided by the fax machines, not by the network. Similar applications are the answering machines or automatic call-center services operating with DTMF signals. In these cases, just like above, the telephone is used for data transmission, while the value-added part of the service is realized by the terminals. The principles according to which the converged global network will be structured will seriously determine the connected business model as well – which is, as yet, significantly different in the case of telephone and Internet networks.

The horizontal design and the end-to-end nature also enables the very fast spreading of Internet applications. The most striking example is WWW that became almost completely dominant on the Internet within 4-5 years. You can never know when a new and explosive application, a so-called ’killer’ application emerges and starts to spread with enormous speed. In the case of napster a totally new application gained ground very rapidly, within just a few months, resulting in very high traffic on the network. Under such circumstances it is rather difficult to design for the future. This fact is one of the important lessons of the Internet: future applications and traffic are very hard to foresee.

However it is not only the applications that come up with new variations, ownership structure of networks and related business conditions are continuously changing as well. The academic network – financed by state funds at the beginning – has been entirely commercialized within a few years. What is more, several ’species’ of Internet service providers were formed, beginning from the ones operating international backbone networks to the enterprises adjacent to

subscribers. Their connections, the exchanging and charging of traffic are continuously changing as technology advances. This change may be accelerated by the emergence of the wireless Internet. What kind of economic frames are ideal for operating such networks? Will the services be provided by several small enterprises connected to busy locations, like airports, hotels or shopping malls via their owners? Or will these be operated by a few large companies? These questions raise not only business, but also regulatory issues concerning interoperation and alliances of networks.

Nowadays Internet is more than a communication network and a source of information, it is also an important stage of the society and the economy. This may result in new phenomena, sometimes problems that we could not imagine before. Information stored and transmitted on the Internet raise copyright issues in an increasing number. The possibility to store electronically, copy and transmit pieces of music urge big music publishers to elaborate new solutions for legal/technological protection, or on the contrary, to agree with Internet-based distributors. The growth of Internet economy and the transmission of important economic data via the Internet make it necessary to develop encrypting and security solutions. At the same time, however, it is required that, if needed, authorities may check the information streaming on the Internet. This inevitably affects legal questions and launches a debate, in which the right to free and confidential communication is opposed to the authorities' right to control.

Internet's international character makes it more difficult to determine legal and economic affiliation of activities on the Internet. Arising from the very nature of the technology, an Internet service provided in one country is accessible in all other countries as well, regardless whether the activity in question is legal in the given country, and if so on what conditions. Furthermore it is not evident which country should account for the results of the Internet economy.

Finally, while Internet use is spreading almost everywhere in the world, we must not forget that this will not reduce the differences between certain countries and regions, what is more, in several cases it will further increase them. Internet

creates an opportunity to get closer to faraway people and lands. The use of the network may make economic and cultural life more efficient, and through this, may give a chance to less developed countries to reduce these differences, but only if the given country can make the best of the opportunity. They have to learn how to use its capacities.

The number of mobile phone and Internet users is rapidly growing in the world. The number of users in itself offers the Mobile Internet, several novel application and technologies the opportunity of spreading. As the saying goes "More is different[1]". An aggregate of cells is more than just a set of cells, it has a new quality, it is an independent living being. Similarly, the joint application of existing and already available technologies in the not so distant future will result something of a different quality, something completely new. If we are able to take the opportunities offered by this enormous change, we will be able to take a huge step ahead.

[1] P. W. Andersson, Science 177; 1972, pp 393-396