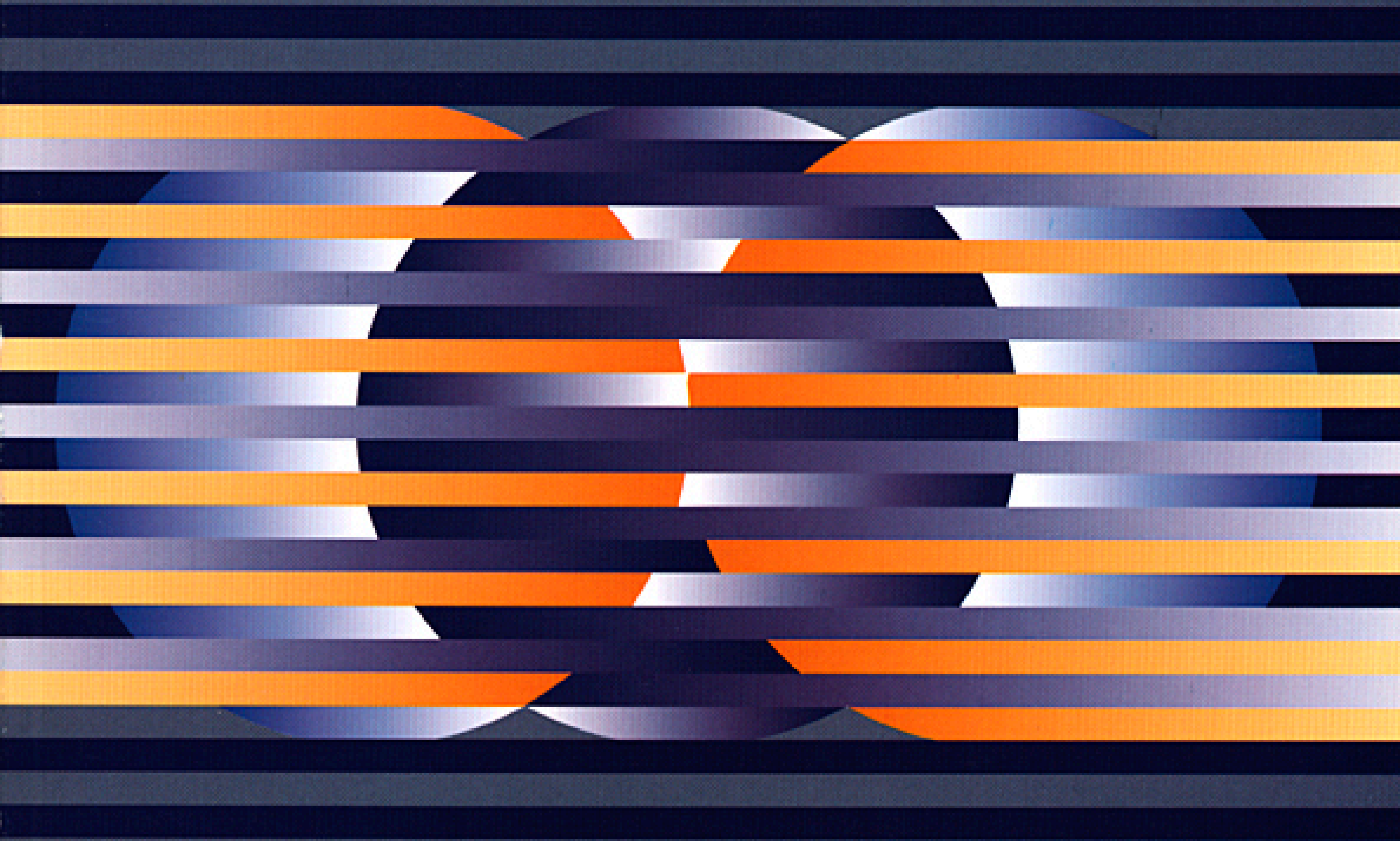
Telektronikk 4.95

Mobile telecommunication



Contents

Feature

Special

Status

Kaleidoscope

Telektronikk

Volume 91 No. 4 - 1995
ISSN 0085-7130

Editor:
Ola Espvik
Tel. + 47 63 84 88 83

Status section editor:
Endre Skolt
Tel. + 47 63 84 87 11

Editorial assistant:
Gunhild Luke
Tel. + 47 63 84 86 52

Telenor

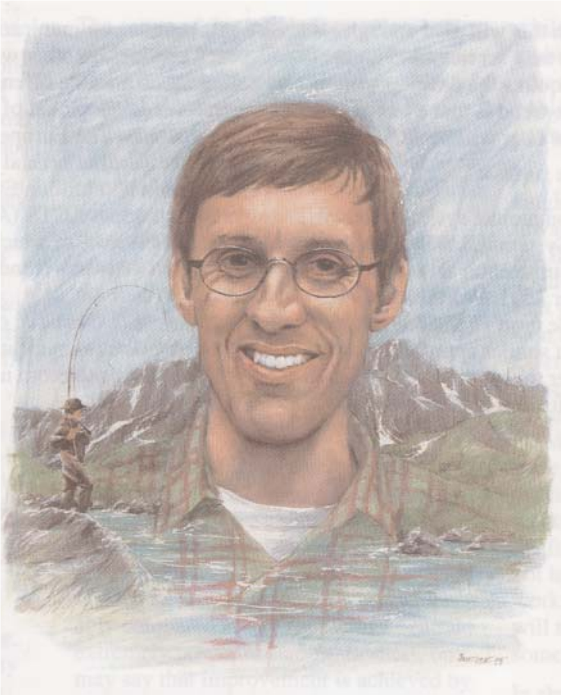
Editorial office:
Telektronikk
Telenor AS, Telenor Research & Development
P.O. Box 83
N-2007 Kjeller, Norway
e-mail: telektronikk@fou.telenor.no

Graphic design:
Design Consult AS

Layout and illustrations:
Gunhild Luke, Britt Kjus, Åse Aardal
Telenor Research & Development

Guest editorial

BY PER HJALMAR LEHNE



Since the launch of the first fully automatic mobile telephone system, the Nordic Mobile Telephone, NMT, in 1981, mobile communications has gradually become the fastest expanding segment of telecommunications business in Norway. NMT was the first system demonstrating mobility and user-friendliness. Before this, public mobile telephony was operated manually, the capacity was low and the telephones heavy (10–20 kg).

Before public mobile systems were established, early users had their own private systems tailored especially to their own needs with a limited coverage and only offering internal communication possibilities. Public systems were established because there was a need for telecommunication services with a greater coverage and an access to the public fixed telephone network.

Among other things, the development of NMT shows that today's telecommunications technology is getting so advanced and complex that it is not possible for a single operator or country to develop new systems due to the cost and time needed. Because of this and the desire of mobility beyond a single country, the directors of the Nordic telecommunications administrations joined in Kabelvåg in 1969 to formally start the development of the NMT-system. At the same time as NMT was launched, the European Post and Telegraph Conference (CEPT) recognised that there was a need for a common European standard for mobile telephony. It took 10 years to develop and standardise this system, now known as GSM – Global System for Mobile Communications. It was when GSM appeared on the scene that the growth virtually exploded, made possible by the digital technology allowing small and cheap handsets, greater capacity and a much greater market than for NMT.

Cordless telephones have existed for many years, giving users mobility, however to a limited extent. It is not possible to use the same telephone at work and at home and when visiting friends. The mobile phone offers a much greater range of mobility than a cordless phone does. Traditional cordless phones may have access to one single base station, thus limiting the mobility of the user, while mobile telephony gives the user mobility over a much larger area. To make this new mobility range possible, mobile communications implies much more than just using radio transmission between the handset and the network. In order to get the system working so that the subscriber actually does not notice the difference between a fixed telephone connection and a mobile phone, mobility functions must be integrated in all parts of the telecommunications network.

New cordless systems, however, like the ETSI-standardised Digital European Cordless Telephone (DECT), will work more and more like a "true" mobile system when it is combined with intelligent services in the fixed network. The fact that traditional mobile systems and cordless telephones in the future will give the user very much the same services will soon be a challenge both to the existing mobile operators running GSM networks and to the fixed network operators.

As the first telecom service to break the monopoly, the paging service was liberalised in Norway in 1989. After political lobbying, the Storting (the Norwegian parliament) decided that two GSM licences should be given to support competition on mobile telephony. The process leading up to this started in 1989 by an application to run a private GSM network in competition with Norwegian Telecom. This triggered a political process. The decision was made to grant two licences on GSM services in Norway. One licence was to be given to Norwegian Telecom (now Telenor Mobil AS, a subsidiary of Telenor AS), while the other was to be given following a round for bidders. This was decided in 1990 and three companies applied for the second licence which was granted to NetCom GSM in 1991. GSM services in Norway were commercially launched in 1993.

The fixed telephone service is still operated as a monopoly in Norway, but all telecommunications services will be liberalised in 1998 at the latest, in accordance with the EU directive. For the large telecom operators still operating in a partial monopoly and partial competitive market, this will undoubtedly be an opportunity as well as a challenge. New operators may enter the fixed network market, but it should also be possible to co-ordinate fixed and mobile services in a way that is not allowed at the moment. Telenor is one of the operators facing this situation.

Where will mobile communications lead us towards and beyond the turn of the millennium? We have probably just seen the beginning. Now mobile pocket phones are not just for the business user or the yuppie. It has become a versatile tool and leisure article for every man and woman. In Norway, there are now more than one million mobile subscribers of which more than 400,000 are GSM, giving the highest penetration in Europe.

This is not without problems. The capacity of today's systems are limited, and GSM will soon be fully loaded in the cities. ETSI soon recognised this to be a problem in the future because of limited frequency resources in the 900 MHz band. A GSM-like system operating in the 1800 MHz-band called DCS 1800 was therefore standardised, giving three times the bandwidth compared to GSM. Therefore, both Norwegian operators are likely to launch a DCS 1800 service in the cities in the near future. Additionally, a third operator has applied for a DCS

1800 license in Norway. But not even this will give enough capacity for tomorrow's demands for new services and higher user bandwidths in the long run. It requires the development of even more sophisticated systems. The trend is that every service offered in the fixed network sooner or later will be demanded on a mobile basis.

In 1991 the work on a third generation mobile system, the Universal Mobile Telecommunications System, UMTS, started in ETSI. A preparatory meeting of the sub-technical committee SMG5 was held in Oslo in November 1991. Work had before this been going on in ITU-R under the term FPLMTS. ETSI/SMG5 started out more or less with a similar approach as when the GSM work started 10 years earlier. Now, the work seems to be shifted more and more towards a further development of GSM.

While mobile communications started out by offering only speech services, in other words, a radio telephone, all communication needs must be offered on a mobile basis in the future. Therefore, development of mobile systems are taking place in several arenas. The European Commission is pushing very hard for the standardisation and development of services and systems. About 630 million ECU is now put into the fourth framework programme ACTS (Advanced Communications Technology and Services) of which Area 4, Mobility and personal communication networks is indicated to take 115 MECU.

The contents of this issue of Telektronikk illustrates the diversity of mobile communications and that mobility is influencing every part of the telecom networks.

The basics of mobile communications

BY KNUT ERIK WALTER AND PER HJALMAR LEHNE

1 Introduction

During the last couple of decades mobile communications has developed from a narrow and specialised field with limited interest to one of the most turbulent and important parts of the telecommunications field. Mobility, in particular achieved through radio based access to telecommunication networks, is one of the hottest issues in the sector today. The concept of mobility is also paving its way into the fixed networks, giving the users flexibility and new opportunities for efficient use of telecommunication.

The liberation of telephone numbering from strict geographical constraints is the major difference between the mobile networks and the traditional fixed networks. This is just the point where recent developments on the fixed network is seen through the Intelligent Network (IN) concepts and thereby making the distance between the two network types smaller. The lack of functions handling mobility in the fixed network is the reason why cellular networks until today have been implemented as separate stand-alone networks.

Mobile radio communication is a broad term, and it covers a variety of systems and techniques if not defined in a more narrow sense. Along the service-axis we have dedicated systems for speech, data, messaging, or simply paging, and we have systems combining a set of services. Radio coverage can be provided by ground based stations or by satellite communication. Further, there are public systems and private networks, different technologies and frequency bands, etc.

Mobile communications is often considered as one specialised field within telecommunications, but in fact it involves a series of different areas:

- Telecommunication services, both traditional and mobile specific
- Radio transmission, i.e. propagation, properties of different frequency bands, antenna design, transmission and reception, modulation, etc.
- Communication protocols for signalling and user-to-user data transfer
- Network architecture, system configuration, call routing
- Network operation and management
- Source coding, e.g. compression techniques for speech or video signals.

The scope of this article is to give a basic overview of the technical aspects of radio based mobility, in particular the public cellular systems. The focus is put on radio transmission and network functionality. Throughout the text the Global System for Mobile communication (GSM), being the most advanced system implemented so far, is used as an example to illustrate the concepts. The intention is not to give a comprehensive description of GSM, but to use the GSM solutions to explain the basics of modern cellular mobile radio networks. Being an ETSI standard, GSM is completely specified in ETSI Technical Specifications – ETS [1].

The basic idea behind a cellular network is that coverage of an area is achieved through a set of base stations, each usable within a smaller area called the cell. By reuse of frequencies the capacity of such a network is increased considerably compared to a design with a central extremely powerful station. In short, one may say that improvement is achieved by adding complexity instead of power. A common way to divide the geographical coverage area into cells is to use a hexagonal pattern as shown in Figure 1.

2 The services of a cellular network

The invention of mobile communication was originally market driven, since there was an apparent need for communication with mobile units, e.g. for different authorities such as police and fire brigades, but also for service providers such as taxis and transport companies.

These early users of mobile communications most often have their own private systems designed especially to suit the applications. Quite often the main traffic demand in these networks is internal, e.g. between a central office and the mobile fleet. The networks offer both traditional services such as telephony and application specific services (e.g. broadcast channels and group calls).

The motivation behind establishment of public systems was that the need for "on the move" telecommunications is not limited to the specific areas above. A market was identified, consisting of users without a need strong enough to justify a private network, or with requirement of mobility over distances beyond what was commercially possible for a private network. For these new users the main demand usually was to get access to the fixed public network, e.g. to be available while travelling. Therefore, the public mobile systems have all the time acted as access networks to the already existing public network, with speech as the dominating application.

Just as we have witnessed a change from single-service dedicated networks towards service integration by ISDN on the fixed network side, the same trend applies to the mobile systems. While the first generation systems provided extension to the PSTN, GSM is intended to extend ISDN services to mobile users. By supporting other services than speech through the digital transmission method, GSM is entering markets which earlier were served by dedicated systems such as wide area paging and mobile data networks. But as on the fixed network side, the first step of service integration does not immediately make the existing networks disappear. The dedicated networks will therefore coexist with GSM for some time.

In short, the services of GSM is constructed as in ISDN: A set of bearer services constitute the mechanism for provision of teleservices. As an example a transparent data connection may be used for facsimile. The bearer services are provided by different formatting and channel coding on the basic physical channel structure, and the teleservice is given by the application. Due to the limitations on the radio interface the highest bitrate supported by GSM is 9.6 kbit/s, a fact that puts some restrictions on the service integration between GSM and ISDN.

Speech is treated specifically in GSM by dedicated codecs. The standard "full-rate" codec works on approximately 13 kbit/s before channel coding is applied. In the mobile station the encoding and decoding may be directly between an

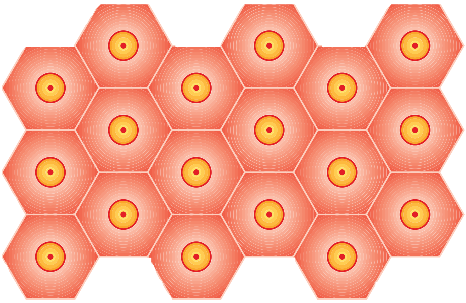


Figure 1 Hexagonal cell pattern

analogue signal and the 13 kbit/s, while a transcoding towards standard 64 kbit/s PCM is required on the network side.

Although subjective testing of the algorithm at an early stage indicated that the speech quality experienced should be at least equal to the best analogue systems, the initial user feedback on GSM was a little bit discouraging on this point. However, it has been shown that most of the problems can be removed by careful design and tuning. Currently, a half-rate codec is on its way, with the potential of doubling the traffic capacity of the networks. Again, there is some uncertainty whether the speech quality performance of this codec is sufficient. Just recently a new and enhanced full-rate codec with substantially better performance has been chosen – this will secure the future life of GSM as a competitive standard worldwide.

In addition to the ISDN-like services GSM also has a messaging capability called Short Message Services (SMS). Two different categories exist: Point-to-point messaging and cell broadcast messaging. In the point-to-point case the messages are addressed and transferred between a mobile terminal and a node called Short Message Service Centre (SMSC). The SMSC is providing interface to external users e.g. through a public data network or PSTN. Mechanisms for storage of messages if the mobile receiver is not available are included, as well as alerting when the mobile is re-entering the network.

The Cell Broadcast Service is unaddressed information sent to all mobiles within a selected cell or group of cells. Typically, this may be used for information relevant to a big number of mobile station, e.g. traffic information.

Altogether the Short Message Service may constitute a basis for development of several new applications, alone or combined with the other services of GSM.

If we look a little bit into the future, the trend is to widen the services offered even further. The GSM system is going through an evolution on the teleservice side, and also new bearer services will be introduced. As an example broadcast, group call and priority functions are defined in order to be able to offer similar services as the private networks.Work is also proceeding on circuit switched data at higher speed as well as a packet radio mechanism enabling a number of users to share a common channel. The work towards the next generation systems is focused on further integration so that the system (or system family) will encompass a wider range of services on the same technological platform. In ETSI terminology the coming system is called UMTS – Universal Mobile Telecommunication System, and the aim is to make a flexible design so that it may encompass more or less everything from rural satellite access to indoor systems with extremely high traffic density. When the fixed broadband networks introduce new bandwidth-consuming services, there will also be a demand for mobile extensions.

Starting out as a professional tool for the few, mobile based services are now just entering the mass market, and in the coming years we will undoubtedly see further developments in the field.

3 Radio communication

The basics for all mobile communication is the use of radiowaves as a carrier for information transfer. A radiowave is an electromagnetic field travelling through a non-conducting media. Although radio transmission is theoretically possible on any frequency, practical limitations such as antenna sizes, the amount of information to be transferred per time unit and the different signal losses at different frequencies reduces the possible usable frequencies for mobile communications considerably. In order to achieve usable radiation the antenna size must be of the same order as the carrier wavelength. Therefore, the practical frequency band for radio communication to and from mobile terminals starts at a few MHz and is currently limited to a few GHz.

There are two main limiting functions for radio transmission:

- the transmit power available in the radio units, and
- the amount of spectrum available for the application.

There is a simple rule of power requirements for a radio connection: the more bandwidth you want to use (in order to increase the information transfer rate), the more power you need. This is quite obvious since a wider bandwidth means collection of more noise power, and hence more signal power is needed to keep a sufficient signal-to-noise figure.

Since the usable radio spectrum is a scarce resource it is subject to rigorous regulations, both nationally and internationally. Certain frequency bands are allocated to different applications and there are strong requirements on the characteristics of the equipment to be used. Unwanted emissions into other frequency bands is one example where such requirements are set. The challenge of any public mobile communication system is to make efficient use of the allocated spectrum, since it directly affects the total amount of traffic to be carried by the system, and thereby the potential income of the operator.

Common for all public mobile systems is a configuration where the radio transmission always is conducted between a fixed site (the base station) and a fleet of mobile stations. A duplex channel of any form is used, and in the case of mobile-to-mobile communication two channels are occupied, even if the mobiles are in the vicinity of each other. This is justified by the simplification on the network side (all connections are treated the same way) and since the amount of traffic where direct communication could be possible is negligible compared to the total traffic.

In private closed radio networks the situation is quite different. Here, the majority of traffic is within the system, and many applications need direct mobile-to-mobile transmission as well as all-to-all communication patterns. The most advanced systems, e.g. used for military applications, are fully distributed so that every mobile station can act as a repeater for communication between other stations. These system types will not be treated any further here.

The configuration of public mobile systems with coverage provided by many base stations has given rise to the name “cellular networks”. The idea is that the coverage of each base station is limited to a certain area (the cell) so that the same frequency resources can be reused in other cells at a sufficient distance. This key point in order to achieve the traffic handling capacity needed with the available resources is taken somewhat further in section 3.6.

3.1 Propagation

The fundamentals of all radio communication systems are the ability of radio waves to transport energy through empty space and other non-conducting media. The antenna is a transformer between bound and free electrical energy. The bound energy being the antenna current and voltage and the free energy being the electromagnetic wave. An isotropic antenna radiates equally in all directions and is commonly used as a reference even though such an antenna does not exist in real life. A practical antenna is directive, i.e. it radiates more in one direction than in the others. In the maximum direction the field becomes stronger than from an isotropic antenna being fed the same power. We say that the antenna has gain, even though it does not comprise any active amplifier. The explanation is that energy is being moved from one direction to another. The total radiation is the same. The "effective isotropic radiated power" (EIRP) is calculated as the power needed from an isotropic antenna to give the same signal strength at the receiver.

For free space transmission the field strength is reduced by the factor of r^2 at a given distance, however, free space is basically not the case in mobile communications. The transmitter and receiver antennas are placed some metres above the ground and the propagation path between them is very complex. Several conditions affect the propagation as shown in Figure 2:

- Parts of the signal are reflected from different surfaces as buildings, hills, mountains, etc.
- Diffraction effects occur when the signal is partly obstructed (by a building or a hill) and through forests
- Precipitation like rain and snow may increase the path loss considerably.

The result is that the signal received is a combination of signals having travelled different paths from the transmitter antenna. In the general case there might be signal components from any direction adding to the total signal detected. If there is line-of-sight between the transmitter and receiver the direct path may be dominating, but this is not always the case. The signals have different amplitude and phase, thus sometimes adding constructively and strengthening the signal energy while sometimes adding destructively whereby the signal is weakened. The consequence of this is that the signal amplitude varies in space and so when the mobile moves around, an amplitude vs. time variation is observed. This effect is called fading.

Exact calculation of the received signal strength is generally impossible due to the large number of different signal paths and effects. Statistical analysis is used in order to predict the signal values. Probability distribution functions are identified to describe the effects.

Fading conditions are often divided into two main categories, short-term and long-term depending on their characteristics and sources. The short-term fading is the previously described case where a small movement of the antenna (or a change in one or more of the signal paths) results in a new value of the received power. The long-term fading is due to a more permanent change of one or more of the signal components, e.g. by shadowing effects from buildings.

Fading only describes the amplitude variation of the signal. A further complication comes from the fact that the set of signals arriving at the receiving antenna have travelled different path lengths. This multipath reception gives a time dispersion of the received signal. As long as the bit duration of the information signal carried by the radiowave is large compared to the typical variation of delay in the received signal, the dispersion causes only minor problems. But when the information rate is increased this picture changes gradually, and severe degradation of the detected signal may occur although the received signal strength is quite high. A common measure for this time dispersion is the "delay spread" of the channel. The delay spread is a measure of the distribution in arrival time for the different signal paths. It is obvious that the significance of the multipath effect increases when the system bandwidth increases. Therefore, the flat fading described first is often called a narrowband model, while multipath giving time dispersion is a broadband phenomenon. Radio network planners are facing this problem today when moving from narrow-band systems such as NMT to GSM with a much higher information bitrate.

As can be seen from this brief description it is not an easy task to predict the coverage area from a cellular base station. One approach is to conduct a set of measurements, categorise the environment (urban, suburban, hilly, etc.) and set up an empirical formula for the received power as a function of the distance. Computer tools are available for predictions of this kind, usually linked to electronic map databases. However, a tool taking into account and calculating delay spread has not yet been introduced.

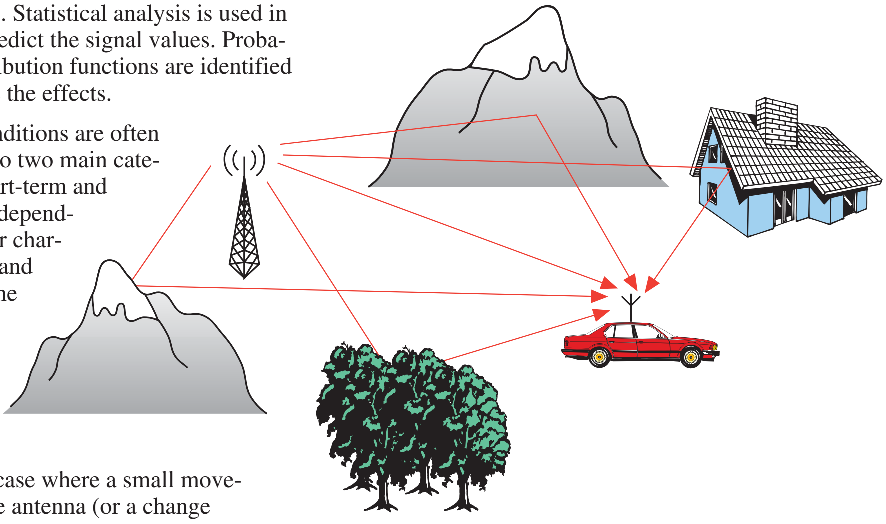


Figure 2 Mobile propagation conditions

3.2 Modulation and demodulation

The signal containing the information to be transferred (speech or data packets) is transponed into the radio frequency band in a process called modulation. A radio carrier is modified by the baseband signal in order to transfer the information content to the receiver.

A modulated radio signal can generally be described by the following equation:

$$C(t) = A(t) \cdot \cos(2\pi \cdot f(t) \cdot t + \varphi(t))$$

Here $A(t)$ is the amplitude (signal strength), $f(t)$ is the frequency and $\varphi(t)$ is the phase. As the equation shows any of these three parameters of the signal can be varied according to the baseband signal. The name of the parameter varied gives the name of the modulation method, hence we get amplitude-, frequency- or phase-modulation. All these methods are used in different radio systems. In Figure 3 the time variation of the signal and the frequency spectrum is shown when a carrier is modulated by a square pulse train for the three basic methods *amplitude shift keying* (ASK), *frequency*

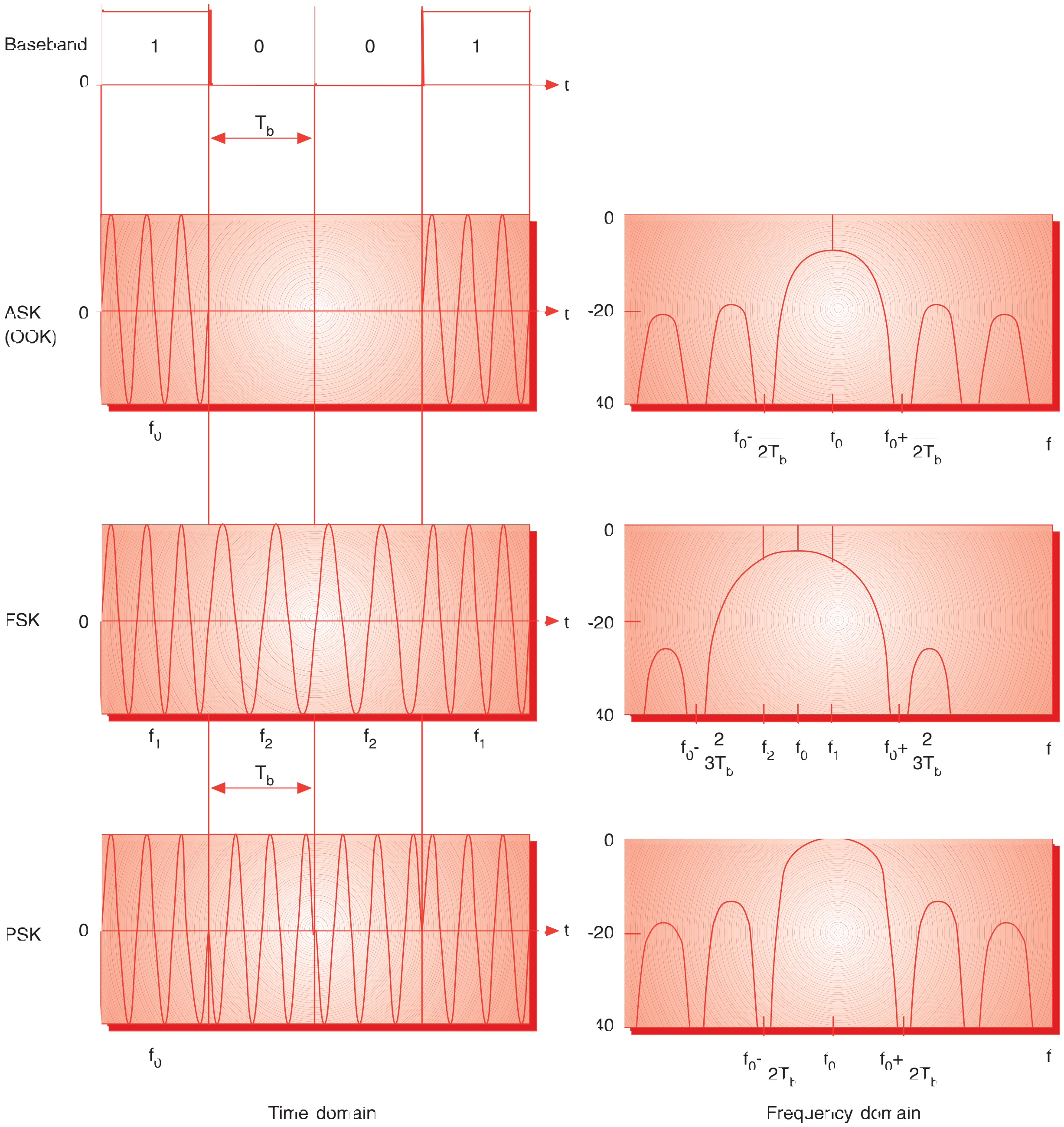


Figure 3 Modulation methods

shift keying (FSK) and *phase shift keying* (PSK). (The word "keying" is often used instead of modulation when the baseband signal is digital.)

The fading pattern of the mobile radio link results in amplitude variation of the received signal. If also the information to be decoded is depending on the amplitude, this will cause noise in the detected signal. Therefore, modulation methods with constant envelope is sought for mobile systems. This rules out amplitude modulation while frequency modulation as used by the analogue first generation systems such as NMT and TACS fulfills this requirement. Phase keying is also giving constant envelope and is often used for digital systems.

A special digital modulation method which is very much used in mobile communications is *minimum shift keying* – MSK. This method can be regarded as a special case both of frequency and phase shift keying. To improve spectrum efficiency, a continuous phase is maintained by using a minimum difference in signalling frequencies. Thus, MSK belongs to the general class of continuous phase frequency shift keying – CPFSK. As other FSK methods this also gives constant envelope making it robust in a fading environment.

Another important requirement is that the bandwidth of the modulated signal must be limited. A digital system using pure phase keying is obviously not a good solution as can be seen from Figure 1. The reason for the broad spectrum is of course the spectral properties of the digital baseband signal (a square pulse has infinite frequency spectrum). By filtering the signal before modulation a much narrower frequency spectrum can be achieved, and the original pulse shape is restored in the receiver. The filter characteristics are giving names to these modulation methods, for instance the one used by GSM is called GMSK – "Gaussian-minimum-shift-keying".

The demodulation process in the receiver transpones the composite radio frequency signal to a baseband signal again. This part of the receiver is a crucial one since it has great impact on the total performance of the communication link. As explained in section 3.1 the mobile radio link may give rise to degradation in terms of delay spread. In GSM this is catered for by the use of an equalisation technique which improves the receiver performance. By utilising a known bit-pattern in the "midamble" of every burst, the receiver is tuned to the specific characteristics of the channel being valid for that burst. The equaliser adjusts the signal before it is decoded by the demodulator. This method is possible by the use of digital signal processors.

3.3 Channel-coding

The purpose of channel-coding is to make the baseband signal better suited to be transferred over the radio path. The basic idea is to introduce some redundancy in the signal which can be used in the receiver to increase the performance, i.e. to reduce the bit-error-rate. The channel code is used either to detect errors or both to detect and correct errors. In the first case the channel code can be combined with an automatic retransmission mechanism (ARQ) to improve the error characteristics of the channel, of course on the expense of longer delay. Normally, this is used for signalling and data services. The second use of channel coding is called FEC – forward error correction, since the corrections are done directly by the receiver. This is useful for all applications including speech since the additional delay is quite small. A detailed explanation of how this works is beyond the scope of this introductory article, interested readers are referred to [1,2].

As described earlier the mobile radio link is characterised by variations in the received signal level and thereby also typical burst-patterns of errors. Unfortunately, the performance of channel codes are best for single errors smoothly spread out in time. To solve this problem the concept of interleaving is introduced. Interleaving is simply to alter the order of bits in the baseband stream so that neighbouring bits are separated. A burst of errors is then spread out by the de-interleaving process in the receiver before the decoding is started. The only drawback of this is the time delay introduced, both on the transmitting and receiving side.

3.4 Multiplexing

By multiplexing we mean the technique used to split the total amount of radio spectrum into smaller units called channels to allow several communication transactions simultaneously. A system is allocated a certain amount of spectrum to be used for communication from the base stations to the mobiles and vice versa. Two different decisions must then be taken: Which method shall be used to distinguish the communication directions (duplexing method) and how shall the different channels be separated (channel multiplexing).

Traditionally, both duplexing and channel multiplexing are accomplished by frequency division.

Separate bands are used for uplink, i.e. mobile-to-base, and downlink (base-to-mobile) communications, they must be separated by a certain frequency distance to avoid disturbance between the two directions. (A so-called duplex-filter is needed in order to stop the transmit signal to get into the receiver part in the radio equipment.)

Frequency division multiplexing (FDM) is the traditional method of getting single channels out of a certain frequency band. Each channel is allocated a certain bandwidth around a centre frequency and the set of channels together fill up the available band.

The most important merits of FDM are the need for high frequency stability (frequency drifting results in disturbance of neighbouring channels), relatively low power requirements (compared to TDM) and no need for synchronisation. NMT, TACS and AMPS are pure FDM-systems with channel spacing 25 or 30 kHz.

By splitting up the time instead of frequency TDM (time division multiplexing) or TDD (time division duplexing) are achieved. Time synchronisation between the mobiles and the base stations are needed in order to avoid disturbance of neighbouring channels. Since only a part of the time is available for transmission, the baseband signal bitrate must be increased compared to FDM. This results further in higher power requirements on the transmitters. On the other hand, it is easier to accomplish different channel types in a TDM scheme, thus making the evolution potential of the system bigger. Another advantage is that the same receiver can be used to monitor the adjacent base station in order to be prepared for handover.

Both schemes FDM and TDM are shown in Figure 4.

GSM uses a combination of FDM and TDM where 8 basic channels are multiplexed by TDM onto every radio carrier, and the carriers are separated in FDM manner with 200 kHz spacing. Figure 4 illustrates the combined division along both the time and frequency axes. Frequency duplexing is used, but the transmitter and receiver timeslot for individual channels are shifted in time so that a mobile station does not transmit and receive at the same time. For a detailed description of the GSM radio multiplexing, see [3].

The last multiplexing method available is a little more complex, namely Code Division Multiplexing (CDM). The idea here is based on the principles of spread spectrum where overlaid codes are separating the different signals transmitted in the

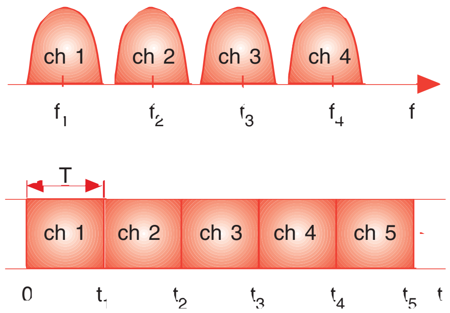


Figure 4 Frequency and time division multiplexing

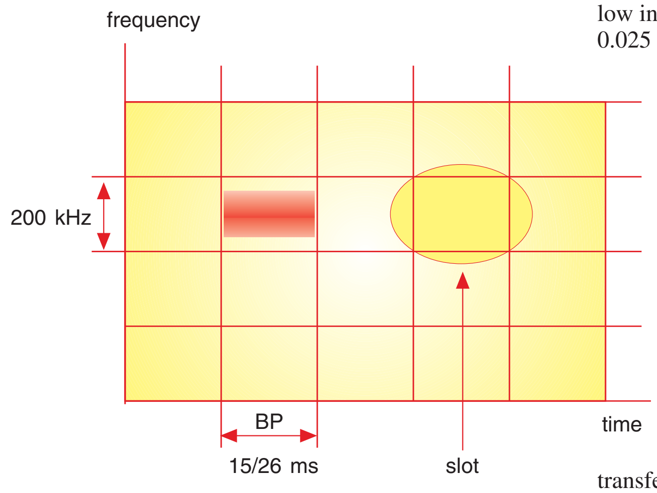


Figure 5 Time and frequency domain division in GSM

same band. The bitrate of the resulting signal is much higher than the baseband signal and the codes are carefully chosen so that the correlation between different codes are small. Hence, each unique code serves as a channel and can be used for communication between the base and a mobile. The transmitter and receiver must be synchronised, and there is another important requirement: quick and reliable power control is needed in order to allow correct decoding of the composite signal. The variation of signal strength received by the base from a mobile close to the cell site and a mobile at the edge of coverage when no power control is used, is higher than the correlation distance enabling the receiver to separate the signals.

CDM used as a multiplexing method is quite new, and there are systems in a starting phase in the USA. The method has impact not only on the channel multiplexing, but several other functions. New behaviour is introduced, instead of a discrete blocking when all channels are used in a cell, a soft degradation can be achieved since the number of codes are high. (More channels used means more noise in each channel.) The amount of equipment installed will of course give blocking sooner or later, but a network using CDM may have better ability to handle variations of traffic.

This subject is comprehensively described in [4].

3.5 Channel access

A cellular system is characterised by a large number of mobile stations but with low intensity of use. Normally, a value of 0.025 Erlang or less per terminal is used when calculating the total amount of traffic in a system. Obviously, there is a need to distribute the communication channels to the mobiles needing them, and use no resources on mobiles in idle or turned off mode.

Since mobile telephony has been the main application of the cellular systems so far, circuit switching has played a major role compared to packet switching since the requirements for real-time transfer and more relaxed set-up times is easier to accomplish by circuit switching. A channel is then assigned to the mobile as long as the call lasts. If the mobile moves out of the cell and into another, a new channel is assigned in the process called handover. There is always an element of random access when a new call is established since the mobile must make the need for channel resources known to the network. In NMT idle marking of traffic channels are used and the mobile picks a channel randomly. In the case of collision the handshake procedure with the network fails and the mobile must make another try.

GSM uses a common signalling channel where uplink timeslots are available for random access from all mobiles. Once a request has come through, the network assigns a dedicated channel for further use (signalling and user-to-user information transfer). The random access protocol used on the access channel is called slotted ALOHA and is also used by some packet radio networks.

In mobile data networks the traffic is of another type and the drawbacks of circuit switching such as long setup delay and low flexibility are becoming serious. The problem is similar to the situation in data networks, and many of the same protocol principles are used. The radio channel has however specific characteristics, e.g. it is impossible for a mobile to decide whether a channel is free or not without the help from the base. The disturbing signal received at the base station may not be present at the mobile site. Therefore, specific radio packet protocols are developed, being slightly different from the ones used for instance in LANs.

Since one of the requirements on next generation systems is to be efficient for heterogeneous traffic, a more sophisticated channel access scheme than circuit switching will be needed. The most promising so far is of the reservation type, where the part of transmission with a risk of collision is kept small and a great flexibility in terms of assigned resources is possible.

3.6 Frequency management and radio network configuration

All frequency use is subject to public control by different regulatory authorities. A specific system gets permission to operate in a given band e.g. in one country. To achieve international roaming the same frequency band must be allocated to the system in all the countries involved. This is what we have achieved in the 900 and 1800 MHz band for GSM/DCS on European basis, and there is good chance for a global band around 1.8 – 2.2 GHz. The international frequency coordination conference WARC (which is a UN-body) has identified bands here but the nations are not obliged to follow this.

It is crucial for a network operator to make efficient use of the available frequencies. As a consequence of this an important property of cellular radio is *frequency reuse*, i.e. the same channel may be used again at a sufficient distance. The challenge is to design the network to match the traffic demand as good as possible. There are several methods to use and there is always room to increase the performance. The following parameters indicate some aspects of this “optimisation” problem:

- the cell site
- the antenna pattern
- the radio channel allocation
- the parameter settings of the cell.

The cell sites must be carefully chosen so that the wanted area and preferably nothing more is covered. In rural areas with small traffic a large coverage area is wanted, and the cell sites are often relatively high. In cities the challenge is to reduce the coverage in order to allow reuse of the same channels not far away. As an additional tool hierarchical cells are often used, e.g. a microcell structure is providing coverage for most of the area, and a so-called umbrella cell is used to cover possible holes in the microcell structure. This overlaid cell must then be marked as “last choice” in order not to be fully occupied.

The antenna pattern is an additional tool for shaping the cells, and makes it possible to have different cells covered from the same base station site.

The interference between different channels is the limiting factor of the cellular systems. The systems are designed such that the most significant degradation factor in high density areas is the interference coming from other users of the same system. In a pure TDM system all cells are using the same carrier frequency, but neighbouring cells cannot use the same timeslot channels. The challenge is then to decide which base stations can use the same timeslot channel while keeping the disturbance beyond the limit. In an FDM system the channels are separated by the frequency, and again the radio channel allocation challenge is to define the reuse distance of the frequencies. Although a theoretical reuse distance in terms of cells can be calculated for a given system, in real life things are not that easy. The different cell sizes, the topography and different traffic demands make this a challenging engineering area where computer tools are available for help.

A modern cellular system also has lots of parameters enabling the network operator to adopt the network to the traffic. The most important ones are parameters controlling the cell selection and handover processes. There is often a conflict between the total performance of the network and local performance within one cell, trade-offs must be taken between local (although severe) problems and potential avalanche effects in bigger parts of the network.

Other examples of parameter settings having impact on the traffic capacity of the network is the discontinuous transmission (DTX) and slow frequency hopping of GSM. DTX used on a speech call means that the output signal is suppressed for approximately 40 % of the time. This has of course great impact on the interference experienced by the others. Slow frequency hopping distributes the effect of strong interference and means that a higher potential level may be tolerated.

Altogether careful design of the radio network is an important success factor for a cellular network operator.

4 Cellular network functions

Although radio transmission is central in mobile communications, great challenges also exist on the network side in order to get a full scale public cellular system working. The requirement that it shall be possible to call every mobile phone from any phone connected to the public network and vice versa means that the cellular systems must interface and connect to the public switching systems, and in particular to the signalling systems. Moving from the traditional situation in telecommunication networks where you could associate every telephone number with a line connected to a terminal and a subscription, to a world where the terminals and users are moving freely around in the network is a huge step from a networking point of view.

As in many other areas of technology, the history of mobile radio has developed from a situation where all the "intelligent" actions were performed by human beings, to a new era where specialised computers take care of everything.

There are two main problem areas that arise when the user with his terminal goes mobile:

- The network must decide where to place incoming calls to the mobile user
- The network must identify a mobile making an outgoing call, e.g. for charging purposes.

Since one basic feature of mobile radio is that no permanent connection between the terminals and the network exists, there will always be an element of searching in the routing process. In the early days the coverage areas of each cell was wide and the number of terminals was limited. In that situation the manual operator could place a calling signal on the air in a large area and the probability that the called user was within reach was high. But each area had its limits, and the calling party often played an important role in order to determine which region to search. The Inmarsat systems still have something left from this situation since the caller must select an "ocean region code" according to the assumed position of the terminal. This is acceptable when most of the terminals are on board ships, and since there are extensive overlapping areas between the regions. In modern cellular radio this mechanism is not sufficient, and automatic processes called "location updating" and "paging" are needed. These procedures are described in more detail later.

Identification is a challenging area in mobile communication. Since all signals are received on the air, the terminals must transfer their identity to the network whenever a communication is started. In the early systems this was done manually, the mobile user had to tell the operator her number (which was a number used both for calling and as basis for charging) before the call setup request was accepted. Modern systems of course deal with this automatically, and verification of the identity is an integrated part of the signalling sequence which is needed in order to protect both the network operators and the users against fraud.

4.1 Network configuration

The cellular networks of today are standalone networks in the sense that mobile-to-mobile calls normally can be handled without using functions of the fixed telephone network. Still they are functionally integrated with the fixed networks by the numbering plans so that calls between fixed and mobile terminals are fully automatic. The future trend is to include mobility functions also in the fixed networks, and the cellular specific parts may then be reduced to being a specific way of accessing the network. The new "cordless" standard DECT together with mobility management e.g. provided by Intelligent Network (IN) functionality can be an early example of this. On the other hand political decisions have so far prevented the integration of mobile switches and exchanges of the fixed networks in many countries. When the cellular network and the fixed network are operated by different companies, the cellular network needs its own control nodes and subscriber databases. An interface between the fixed and cellular networks on exchange basis is therefore the normal solution.

Generally, two main types of network nodes exist in a cellular network, *mobile exchanges* (or switches) handling all switching and call control functions, and *base stations* managing radio connections to the mobile terminals. The switch has access to storage functions, either internal or in external stand-alone nodes, where the subscriber data are kept. The ever changing subscriber data is an important point where a mobile switch differs from a traditional exchange in the fixed network.

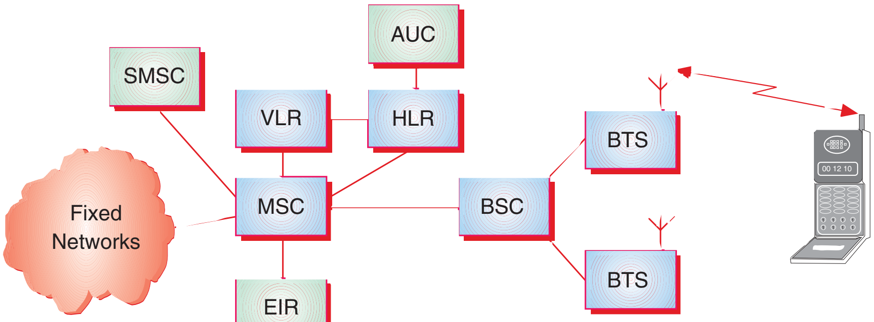


Figure 6 GSM Network architecture

Since the GSM system represents the most advanced cellular network architecture of today, it is shown as an example here, and described a little more in detail.

Figure 6 shows the different nodes which are connected by defined interfaces.

MSC (Mobile Services switching Centre): This is the mobile switch in charge of connecting all calls to and from mobile stations within a certain area. The MSC is connected to the Base Station Subsystem on the "mobile side" and to other MSCs and nodes in the PSTN or ISDN on the "fixed" side. An MSC may be compared to a local exchange since it directly interfaces to the mobile terminals on call level, but it also acts as an intermediate exchange e.g. in the case of roaming users when the call is forwarded to the current position of the user. The MSC is most often co-located and integrated with the local database, the VLR.

VLR (Visitor Location Register): The VLR is a local database containing all relevant data for the mobile terminals currently residing within the area controlled by the MSC/VLR. The VLR is connected to other VLRs and HLRs through the signalling system. In the early GSM specification an interface between MSC and VLR was also defined, but all equipment vendors have built the MSC/VLR as one unit.

HLR (Home Location Register): This is the database where subscription information is kept permanently. Each GSM network must have at least one HLR, and each GSM subscriber is registered in only one HLR. In addition to the permanent subscription information, the location information in HLR is continuously updated according to the roaming of the mobiles. Basically, this location information is identifying the MSC/VLR where the mobile station currently may be reached. The HLR is responsible for the security functions for its subscribers, by managing sets of parameters. All mobile terminated call setups include a request for routing information from the HLR. The HLR is connected to VLRs and MSCs of all co-operating networks by the signalling networks. The generation of security parameters is performed by a "security node" called Authentication Centre (AUC). This functionality may be integrated in the HLR or implemented as a separate node.

BSC (Base Station Controller): The prime function of this node is to perform the functions managing the radio connections towards the mobile users. This includes interconnection for signalling and traffic channels. BSCs of different sizes are available, controlling a number of Base Transceiver Stations from 10 or 20 up to 100 or more. The BSC is connected to the MSC at the so-called A-interface which is standardised in order to allow network operators to have different suppliers for the two node types. Trunking effects are achieved in the BSC since the signals from several BTSs are concentrated towards the MSC.

BTS (Base Transceiver Station): This is the network unit providing direct radio connections to the mobile stations. It basically consists of the radio transmission equipment and functions for connecting the radio channels to the channel structure on the fixed connection to the BSC. A BTS is configured according to the geographical size and traffic load in the cell it serves. The interface between the BSC and the BTSs is called "Abis" and is also intended to be an open one. However, lack of agreed specifications for operation and maintenance has so far prevented this in practice.

EIR (Equipment Identity Register): This is an additional node keeping track of the mobile equipment used. Every type approved mobile station has a unique serial number, and a list of "allowed equipment" is maintained by the co-operation group of operators. The EIR also makes it possible to blacklist mobiles causing problems or if access to the network should be prevented for any other reason.

SMSC (Short Message Service Centre): The role of this external node is to provide the short message service. Several possibilities for input and output of messages to or from mobiles exist, including interworking with electronic mail systems.

4.2 Signalling functions

Signalling functions are the key for realisation of all network functions. A signalling protocol is simply a set of rules for exchange of information between network nodes, and can be compared with a language. If the nodes use different languages, misunderstandings occur, and the wanted result of information transfer is lost.

Again GSM is taken as an illustrative example, being the most advanced system so far also in this aspect. Traditionally, one can distinguish between two different types of signalling: the user-to-network (or subscriber) signalling and the internal network signalling. The user-to-network signalling of GSM is of course taking place over a radio connection and includes functions for securing the connections in addition to the standard call handling functions needed in all telecommunication networks. The OSI principles of protocol layering are used, and the GSM user to network interface is strongly inspired by the ISDN subscriber line signalling. The "intelligent parts" are defined on layer 3 and are grouped into Radio Resource Management (RR), Mobility Management (MM) and Connection Management (CM).

The Connection Management contains functions for Call Control (almost identical to ISDN), Supplementary Services Control and Short Message Service Control. The layout of the air interface protocol stack is shown in Figure 7. Here, also the lower layers for reliable message transfer and physical channel control are indicated.

On the network side most of the GSM internal signalling is based on ITU Signalling System No 7 (SS7). The only exception is the Abis interface where a modified version of the ISDN subscriber line ("D-channel") signalling is used. In the rest of the network the lower-layer mechanisms of SS7, Message Transfer Part (MTP) and Signalling Connection Control Part (SCCP), are providing transfer of information between physically distant nodes. These protocols are commonly used in the fixed networks, e.g. by the Telephone User Part (TUP) and ISDN User Part (ISUP). For internal GSM communication, i.e. between the MSCs and between the location registers VLR and HLR an entire new protocol called Mobile Application Part (MAP) is defined. This protocol supports all necessary exchange of information between these nodes, often on an international interface. The MAP uses an additional end-to-end transport mechanism of SS7 called Transaction Capability – TCAP. In Figure 8 are shown the protocols used between an MSC and a distant HLR.

On the interface between the MSC and the BSC another GSM-specific protocol is used above MTP/SCCP. It contains functions for controlling the BSC operation as well as transparent transfer of messages to the mobile stations.

For call setup to and from the fixed network, the protocols to be used are depending on the local network. Most often TUP or ISUP of SS7 are used, but also other systems like the older R2 signalling system are possible.

4.3 Numbers and identities

In the first generation of cellular systems the association of a number was moved from a line to the terminal, so that every terminal had its given identity serving both as identification and as calling number. Certain requirements exist for the structure of a calling number, since its format must be according to the international telephone numbering plans. The solution was that dedicated "area codes" were selected in each country identifying one or several mobile networks, and maybe also some geographical distinction within the country. The NMT system was the first one specified and put into operation with international roaming. This made it possible to use the same single access number to call a mobile user wherever he might be within the Nordic countries. The price paid for this feature in the early implementations is non-optimised routing, which is visualised in an example later on.

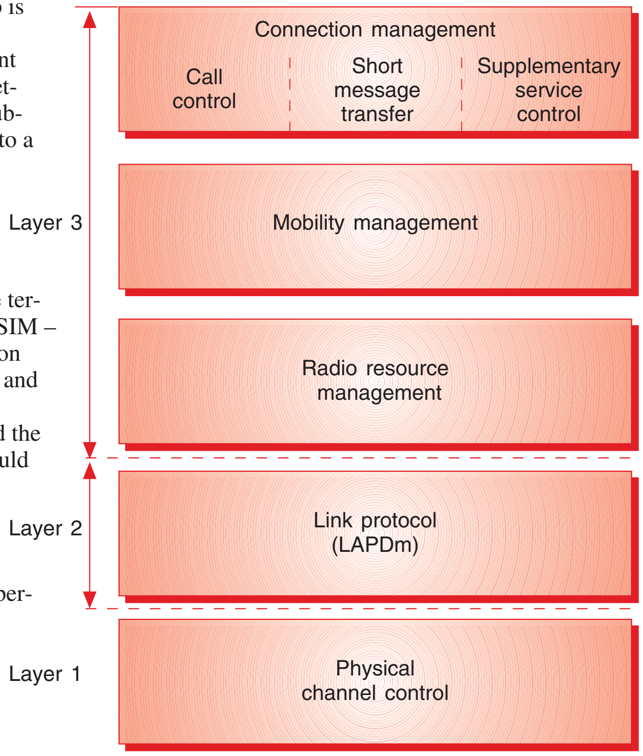


Figure 7 GSM air interface protocol stack

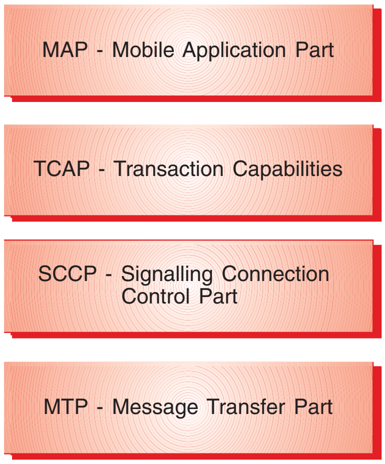

MAP - Mobile Application Part
TCAP - Transaction Capabilities
SCCP - Signalling Connection Control Part
MTP - Message Transfer Part

Figure 8 Protocols between MSC and HLR

This first mobility step is often called terminal mobility, since the point of attachment to the network is free, but the subscription is connected to a terminal. A further evolution is achieved in GSM where the subscription identification is moved from the terminal to a card called SIM – Subscriber Identification Module. Now the user and the subscription are uniquely identified and the point of attachment could be any GSM terminal in any of the co-operating GSM networks. In fact, this is what is called "personal mobility", however limited to a certain environment, namely GSM.

Another step also taken in GSM is the distinction between the directory number (used for calling to a mobile user) and the subscription number (used to identify the SIM). This is mainly done for practical purposes since it allows flexibility when allocating numbers and when renewal of SIM is needed. But it also opens up for new features for instance the possibility of having more than one directory number associated to the same subscription.

Although the subscription identity is all that is needed to enable the GSM service, it became clear that identification of the mobile terminals themselves was needed also in GSM in order to secure the type approval of the equipment. An additional database is defined for this purpose, and the network operators are free to include checking of the equipment identity in the various signalling transactions with mobile terminals.

4.4 Location management

Location management is the name of the functionality used by a cellular network to keep track of its users. The reason for having this function is to avoid a time- and resource-consuming search over a wide service area for every incoming call to the mobile user. The basic principle is that the service area is divided into smaller parts and that the network stores information of the current "position" of every user. When an incoming call is present, the stored information is used to route the call to the current location of the called user. Every mobile terminated call therefore includes a translation from the called directory number to a routing number usable to direct the call to its final destination. The "mobile number" will only identify where to ask for further routing information.

Different possibilities exist for keeping the stored location information updated according to the movement of the terminals and users. First it has to be decided on which level of accuracy the information should be. The term "location area" is commonly used for the area in which the user may move around without any change in the stored location data. The size and limits of the location area may vary from system to system.

Typically, a location area will encompass a number of cells, depending on the traffic load in the area. In the early phase of a network roll out there will often be few subscribers and a small number of switches covering large land areas, thus there is no reason for splitting it up into lots of location areas. In later, more mature states the traffic load is increasing and additional location areas are needed in order to reduce the paging and access channel load.

The most common strategy for keeping the location information updated is to put the burden of information on the mobiles. By reading broadcast information when moving from cell to cell, a mobile will detect when a new location area is entered. In this case it has to access the network and inform that it is present in the new area. This procedure is fully automatic and invisible to the user.

In GSM two additional features are included in order to enhance the service. The first one is called attach/detach and is simply a mechanism that forces the mobile to tell the network when it is powered down, and when it once again is turned on. One advantage of this is that the network can act immediately when a call is placed to a mobile switched off, without losing time and resources on paging. Another advantage is that it enables the system to transfer stored messages as soon as the mobile terminal is turned on.

The other function is called periodic location updating. The original idea behind it is that in spite of quite reliable procedures, there is always a probability for mismatch in the status of the mobile and the network. The mobile might be in the "updated" state while the information in the network is pointing out a different area or no area at all. The result is that the mobile is not reachable even although it is within the coverage area. Therefore, a timer in the mobile station (with its value set by the network) is supervising the idle time of the mobile and triggers a location update procedure when it expires. This feature contributes to the value of the short message service by trigging new attempts of unsuccessful message transfers (e.g. if the user was temporarily without coverage). The mechanism also helps restoring the network's location data after a possible data loss.

4.5 Paging

In a cellular network paging is a procedure to locate a specific mobile for some reason, e.g. setup of an incoming call. It consists of a broadcast message sent out in some area identifying the mobile wanted. Normally, paging is conducted in all cells within the location area determined by the location management function. Although different strategies for this are identified, the most common is to send paging signals in all the cells simultaneously. Usually, the number of paging trials and the time between them can be set by adjustable parameters. When the mobile receives a paging signal, it may access the network in a selected cell and the paging process is completed.

4.6 Handover

One of the most important and specialised functions of a cellular network is the ability to transfer calls from one cell to another when the mobile is moving. In the early days it was called "hand-off", now the term "handover" is used reflecting that the procedures are being more reliable and that the call is "handed over" from one cell to another.

The prime reason for handover is the limited coverage of each cell. The basic mechanism ensures that the mobile is connected to the base station providing the best connection.

There might also be cases where the network decides to move calls even to a slightly worse channel, e.g. in order to resolve overload situations in certain cells.

Depending on the configuration, a handover may involve a number of nodes in a network. As an example GSM has three different cases for handover:

1 Handover between BTSs connected to the same BSC, so-called *intra-BSC inter-BTS* handover

2 Handover between BTSs connected to different BSCs belonging to one MSC, so-called *inter-BSC* handover

3 Handover between BSSs (BTS+BSC) of different MSCs, so-called *inter-MSC* handover.

The handover process can be divided into a preparation phase, a selection phase and an execution phase:

The *preparation* phase consists of information collection in order to enable a correct decision for handover. The systems are quite different on this stage: According to the original NMT specifications nothing is done before the quality of the connection is below a certain limit. When this limit is reached (either on uplink or downlink) the network activates measurements on the base station that it thinks may be candidates for handover. The mobile is not included at all in this process. In GSM measurements are conducted all the time, both on uplink and downlink. In addition, the mobile is measuring alternative base-stations and transfers all its measurements to the network for analysis. The network is responsible for recognising the need for handover based on this input data.

When the need for handover is detected, the responsibility of *selecting* the new base station lies on the network both for NMT and GSM. In GSM the equipment manufactures have freedom in implementing an algorithm detecting the need and selecting the candidate. The new base station must allocate the resources needed before the execution phase can start.

In the *execution* phase information of the new channel is passed to the mobile, and shifting is performed. On the network side actions must be taken in order to minimise the disruption experienced by the user. Since there is always a chance of failure, the old channel is usually kept open so the mobile can return if something goes wrong.

In the case of handover between MSCs in GSM (inter-MSC handover), the original MSC is keeping control of the call since it has all the data e.g. for charging. The original MSC is called the *anchor* MSC. In the case of subsequent handovers, i.e. a BTS connected to a third MSC is serving the mobile, the second MSC is released, and a new connection is established between the anchor MSC and the new one.

While handover is primarily a necessity from the cellular structure, it is also possible to use the mechanism for traffic reasons. If one cell is in trouble due to overload, some of the mobiles may be moved to alternative cells, even if their connections are degraded a little bit. It is, however, not an easy task to design a reliable algorithm for this use.

Since handover is one of the key areas of cellular radio, it is also being heavily studied for coming system generations. A trend is to distribute the control functions more, and to put more responsibility in the mobile, being the one knowing the local radio environment best. The words “seamless” or “soft” handover are used for future systems, both showing that the intention is to hide the process even better for the users. Already in DECT, where all phases of handover are controlled by the mobile, a seamless handover is achieved since the handset may establish a new connection before the old one is released. Another example is a CDM-system where the mobile may keep the same code while moving into a new cell, and the network may use advanced combination techniques to slowly shift the communication from one cell to another.

4.7 Incoming call routing

Specific functionality must be invoked in order to route an incoming call to a mobile since the cellular network supports a single calling number irrespective of the whereabouts of the called mobile.

Figure 9 shows this basic operations of an incoming call, including the following steps:

1 The fixed public network recognises that the called number is a mobile number, and routes the call to a mobile switch.

2 The mobile switch interrogates the location database in order to retrieve routing information.

3 A “roaming number” enabling routing according to the current position of the called mobile is returned.

4 The call is routed to the correct mobile switch, often through the fixed network.

5 The mobile switch controlling the current area is searching for the called mobile, and the call is established in a cell depending on the actual position of the mobile.

Note that the calling user does not need to have any knowledge of the current location of the called user.

A critical point in the routing performance context is at which point the call is recognised to be a special call with a need for interrogation of routing information.

In the public telephone networks of today the only generally available way to recognise a call being of this category is analysis of the called number and routing to a mobile switch. Unfortunately, this will often lead to so-called tromboning especially in the case of international roaming users. Consider the example of a Norwegian Telenor Mobil GSM user roaming in a cellular network in France. A person in France dialling + 47 900 xxxxx would get a circuit to an international exchange in Norway (since the French network has no knowledge of the numbering structure behind the Norwegian country code), then to a Telenor Mobil MSC and back the same way to the MSC in France servicing the area in which the mobile resides. Thus, the call occupies two international links while a short signalling interrogation would have been all that was needed in addition to the local call. This is a very good example to illustrate that although a function may look rather simple, limitations in the existing networks and their signalling systems often make implementation cumbersome.

In the future this inefficiency will be overcome when mobility is included also

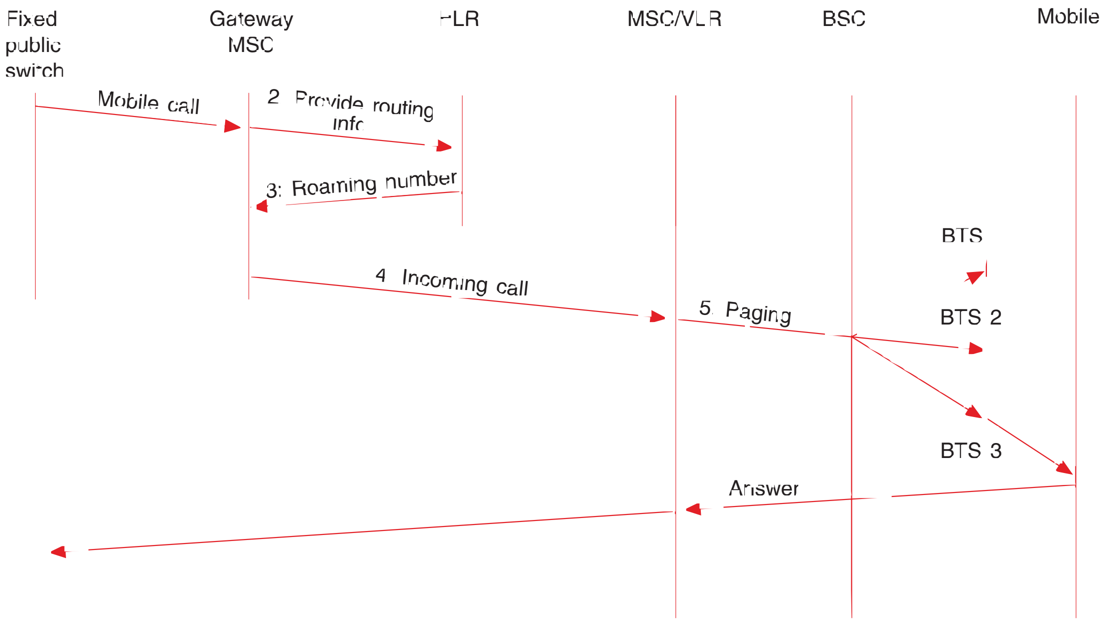


Figure 9 Basic routing of incoming calls

in the fixed networks. Look-ahead procedures in different forms will be implemented.

A step in this direction is taken by the development of Intelligent Network (IN)-concepts. In fact, one of the most central part of IN is the ability to translate a directory number to a routing number. A "normal" call setup will then include at least one signalling transaction providing instructions for the call handling process. The interrogation must be invoked as early as possible in order to achieve the most efficient routing.

4.8 Outgoing call

Compared to incoming calls, outgoing calls in cellular networks are relatively simple from a networking point of view. When a mobile is registered in a location area, outgoing call setup is performed without any interaction with the HLR, since all relevant data has been downloaded to the VLR. The MSC is analysing the called number and initiates routing accordingly. It also performs the charging information collection for later transfer to the home network of the calling subscriber. It is therefore of utmost importance that the mobile is identified correctly before an outgoing call is accepted by a foreign network.

4.9 Security functions

The same features and properties that make radio communication attractive, unfortunately also make it susceptible to fraud and eavesdropping. The first generation systems were almost without any protection in this direction. The analogue FM-modulated signals can easily be detected and the initial lack of verification of identities created a market for illegal equipment, especially in some countries. Fraudulent users represent a great problem both for the network operator and the users. The subscribers receive bills which are not in line with the usage of their mobile, and the network operator loses money when these bills are not paid. Authentication has therefore been added to most analogue systems at later stages.

Under the development phase of GSM the security requirements were recognised, and a set of precautions were specified:

- It should not be possible to listen to a conversation by eavesdropping
- Fraudulent use should be excluded by strong authentication
- It should not be possible to identify the users by listening on the signals on the air.

The chosen solution relies on a secret key linked to each subscription, and one-way algorithms for calculation of ciphering keys (stream ciphering) and signature for identity verification. Use of temporary identities when in un-ciphered mode prevents revealing the real identities in almost all cases.

It is generally assumed that the demand for security will increase in the future for all kinds of telecommunication, the mobile world is for the time being ahead of the fixed networks in this area.

4.10 Charging principles

Charging is an important task of every public telecommunication network, and different charging structures is an area for competition between network operators. Traditionally, the subscribers' bills are composed of a fixed subscription fee and a part directly dependent on call time and destination. The normal practice in the telephone network is that the calling party is charged for the whole connection, and there is often a direct relationship between the call distance and the charge. It is also possible for a calling user to know the cost of a call in advance simply by analysing the called number. Since the cellular networks are functionally integrated with PSTN and ISDN the same principles must be taken into account, at least for mobile terminated calls.

A user calling a mobile user does not know where the mobile is, and can therefore not see from the called number what the real cost of the call will be. From the start the NMT network was limited to the four Nordic countries and it was decided to use a flat tariff for mobile terminated calls. The risk when using flat tariff is of course user dissatisfaction with the charging of what they consider to be local calls. When the roaming possibilities increase and the difference in actual call distance becomes larger, it is obvious that flat tariff cannot be maintained. The solution agreed is that the mobile subscriber must pay for the international extension of incoming calls when roaming. This charging of incoming calls is a totally new concept and additional features for screening, conditional forwarding, etc. are included in order to maintain the service also for those who do not want to pay for receiving all kinds of calls.

For outgoing calls the situation is simpler: charging is accomplished by the serving MSC and all possibilities exist for differentiation according to the called number. The challenge here is more on the administrative side, since the charging information must be transferred to the home network of the calling mobile. Reliable mechanisms are established to make this interaction between the operators possible.

5 Summary

In this article an overview of the most important technical aspects in modern cellular mobile communications has been given. Although being rather brief in most areas, it has hopefully given the reader an idea of the complexity of the systems. The area is an extremely challenging one for all the parties involved, the user equipment manufacturers, the network infrastructure vendors, the network operators, the service providers, and the users. In this fast growing field things quickly get old-fashioned, however, at least some of the material of this article will be valid also for the coming systems.

6 References

1 European Telecommunications Standards Institute (ETSI). *The GSM technical standards.* Sophia Antipolis.

2 Lin, S, Costello, D J. *Error control coding : fundamentals and applications.* N.J., Prentice-Hall, 1983. ISBN 0-13-283796-X.

3 Mouly, M, Pautet, M-B. *The GSM system for mobile communications.* Palaiseau, Mouly & Pautet, 1992. ISBN 2-9507190-0-7.

4 Eriksen, J, Svebak, O D. Code division multiple access : hot topic in mobile communications. *Telektronikk,* 91(4), 99–108, 1995. (This issue.)

The history of mobile communications in Norway

BY LAVRANS GRIMSTVEIT AND HANS MYHRE

This article briefly describes the development in Norway of mobile communications, which started with maritime radio communication in 1908 and was later extended to include communication with aircraft and land mobile units. The advances in this field have been enormous and are proceeding towards personal communication. Norway has in many instances been in the forefront within mobile communication. Examples are maritime communication, mobile satellite communication and cellular telephone services.

The achievements in mobile communications in Norway, particularly during the last 20 years, are to a large extent due to team-work between the telecommunications operators in the Nordic countries, complemented with discussions with equipment manufacturers. This is also described in the article.

Maritime communication

Maritime communication by means of Morse telegraphy was one of the earliest applications of radio communication. Internationally, the first ships were equipped with radiotelegraphy stations around the year 1900. Radio communication was initially intended as a safety measure and for the management of the ship. Later, radiotelephony was introduced and gradually became the most important way of communication, primarily for the purpose of public correspondence.

The public maritime mobile service in Norway started in 1908, when Sørvågen radio in North Norway opened MF radiotelegraphy with ships. Successively, coast radio stations were built to cover the coastal areas of the Norwegian mainland from the Skagerrak waterways towards Denmark in the south to the Russian waterways in the north, a distance of 2650 kilometres, as well as the areas around Spitsbergen, Bear Island and Jan Mayen. It was not until 1931 that MF radiotelephony was introduced in the maritime mobile service in Norway.

After the second world war, radiotelephony gradually took over most of the radio communication with ships. In 1955, Norway had 27 coast radio stations, 12 were equipped with radiotelephony and 15 were equipped with both radiotelegraphy and radiotelephony. In the late 1950s, VHF radiotelephony was introduced for short distance communication with ships. Gradually, a larger proportion of the traffic was carried on VHF. The telephone calls were manually connected through the telephone network.

Long distance communication with ships was made possible in 1927 when an HF radiotelegraphy station was put in operation at Bergen radio. In 1949 the station was also equipped for HF radiotelephony. Bergen Radio had been damaged during the war and had been provisionally rebuilt. The station was unable to cope with the increasing traffic, and it was therefore decided to build a new maritime HF station near Stavanger. The new station, named Rogaland Radio was opened in 1960, and soon proved to be one of the most efficient and busiest HF stations in the world. In 1965 a radio telex service was also put into operation.

Land mobile communication

After the second World War, land mobile communication was introduced in Norway, first in the form of private mobile radio networks. Gradually, enterprises and public utilities recognised the benefit of swift communication with their vehicles, and many private mobile radio networks were put in operation.

There was, however, a strong wish on the part of many users to increase the coverage area, which could most easily be done by installing the base stations on mountains and hills. In this regard, existing broadcasting and radio relay stations were the most attractive locations. Norwegian Telecom, which issued licences for these mobile radio networks, judged a development along these lines to result in a costly infrastructure consisting of several parallel networks with similar coverage. In addition, it would result in poor utilisation of the radio frequencies. Norwegian Telecom also saw a market coming up for radiotelephony for private persons. All these objectives could best be met by a public radiotelephone service.

Manual radiotelephone

In 1966 Norwegian Telecom decided to open a manual public (VHF) land mobile radiotelephone service. Equipment for six base stations were procured and installed according to customers' demand. The service started with 3 channels in Oslo and a small number of channels in other densely populated parts of the country. Initially, the mobile telephone subscribers had to listen continuously for calls on the calling channels. It was a great relief, therefore, when in 1971 selective calling was introduced.

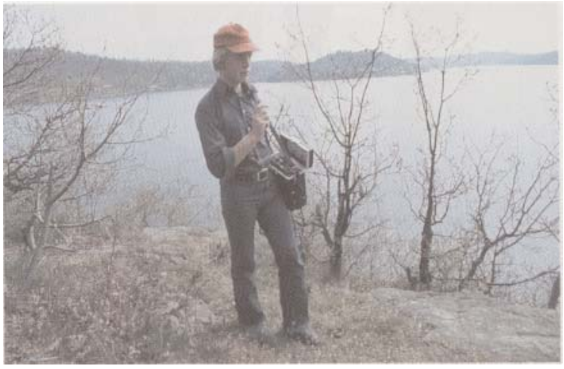
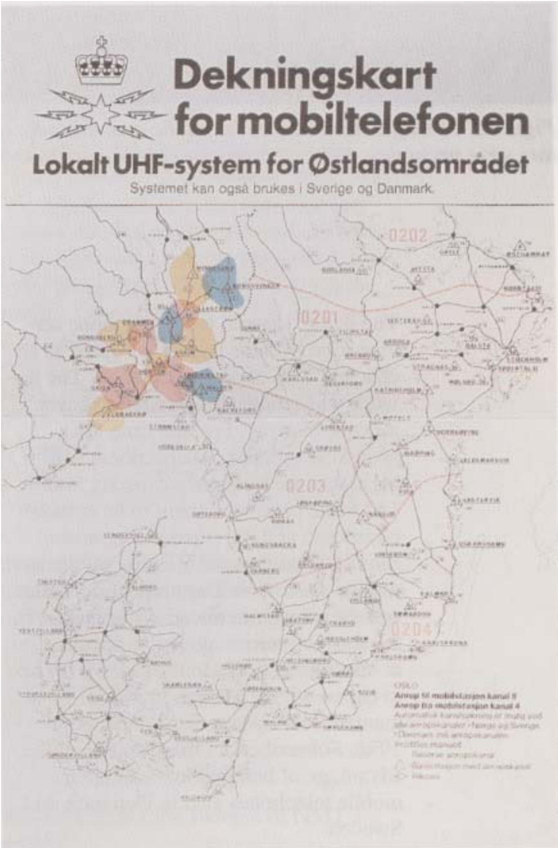


Figure 1 Despite the weight of the equipment, the public manual radio telephone was a great enhancement in making public telecommunications available outside the fixed network



Dekningskart
for mobiltelefonen
Lokalt UHF-system for Østlandsområdet
Systemet kan også brukes i Sverige og Danmark.
0202
0201
0203
0204

Figure 2 The manual radio telephone service covered most of Norway in addition to the central parts around Oslo, as shown here

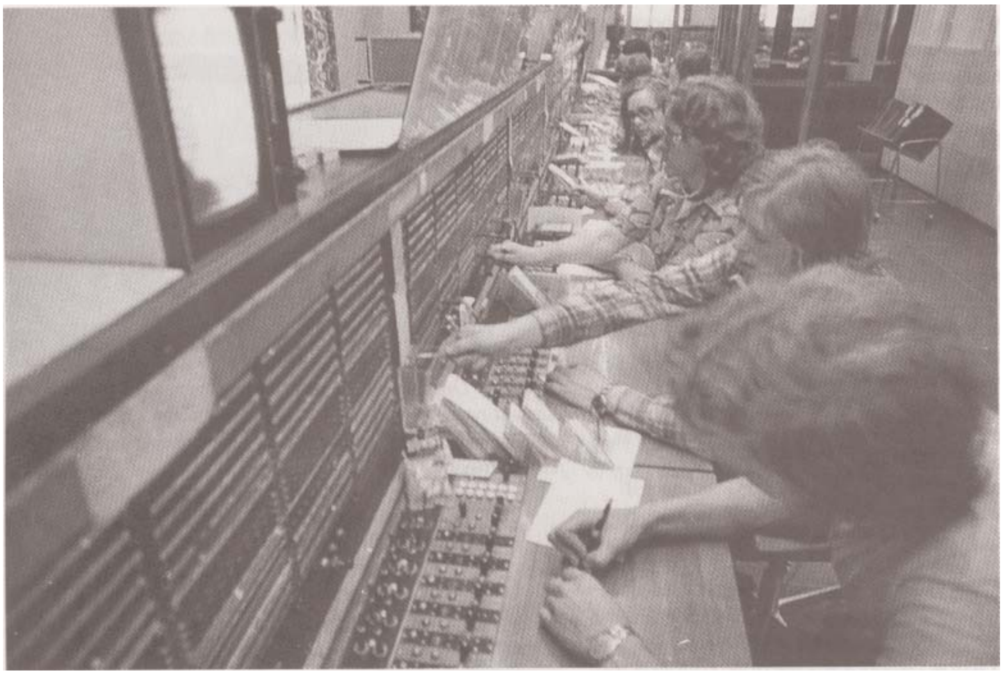


Figure 3 As traffic increased in the public manual radio telephone service, a lot of switchboard operators were necessary. Here is shown the switchboard in Oslo in 1974

The public manual radio telephone service soon became widely accepted as an efficient way of communication. The network was gradually extended to cover most of Norway. For that time the growth in number of subscribers was high, and the frequency capacity in the 160 MHz band was about to be exhausted in Oslo and the surrounding areas. Similar situations arose in the public land mobile systems in Denmark and Sweden, and the telecommunications operators in the three countries agreed to open regional manual networks in the 460 MHz band to cope with the growth in traffic. This manual UHF system was launched in 1976. Subscribers in the system had the advantage of being able to use their mobile telephones also in Denmark and Sweden.

In Norway, the UHF system was built to cover the most densely populated south-eastern area. It was closed down in 1986, primarily because the frequencies had to be used for NMT. The manual VHF network was closed down in 1990. The number of subscribers in the two manual radio telephone systems reached 31,000 at its maximum in 1981, the year when NMT was launched.

NMT

The history of NMT goes back to midsummer 1969, when the Nordic telecommunications directors at their meeting in Kabelvåg in northern Norway decided to establish a working group with the following main task:

> "To study the possibility for future establishment of compatible Nordic public mobile telephone systems, preferably as one common system. This system should probably be automatic, but manual systems cannot be excluded. In any event, it is necessary that the system structure is investigated and a common frequency plan made."

In the beginning, the main task for the Nordic NMT Committee was to find a common frequency band, and to select the vital system parameters. Specific development projects were assigned to the industry to make use of their expertise. An automatic mobile telephone system was designed based on these investigations and findings.

In order to verify the functionality of the design and to adjust the system parameters, a trial network was established in Stockholm. This network only consisted of a "home-built" switch controlled by an industrial processor, and three base stations, each having two channels. Nevertheless, extensive tests were made with four mobile stations which were developed for the testing.

When the time had come to order mobile telephone switches and base stations, the Nordic operators co-ordinated the commercial activities towards the manufacturers, both in the tender process, the purchase and the follow-up of the deliveries.

A main design requirement was that the mobile telephones should be as simple as possible, while the complex functions should be located in the centralised equipment. This would reduce the total system cost. The mobile telephone industry was therefore consulted several times during the design work.

Another reason for listening to the industry's voice was that the market for mobile telephones should be open. A suitable specification would increase the manufacturers' interest in making mobile telephones, reduce the manufacturing cost, and thereby reduce the price for the end user. This proved to be successful with regard to cost as well as weight of the equipment.

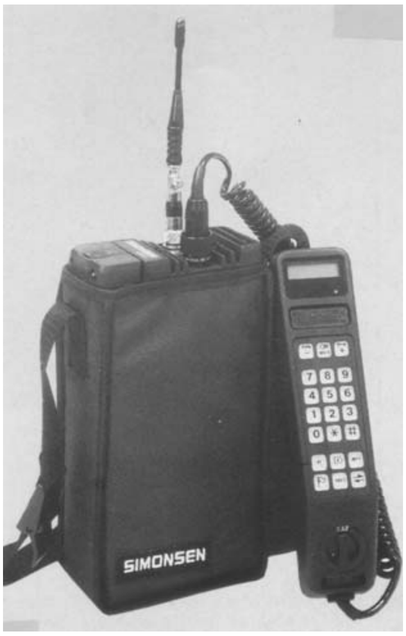


Figure 4 The first automatic mobile telephones of the NMT system were more portable than "handy" unless installed in a car. The picture shows an early NMT-900 from the Norwegian manufacturer Simonsen

Oslogryta
koker over.
Antall samtaler
200
150
100
50
Fra mobil
Til mobil
Klokkeslett
06
08
10
12
14
16
18
20
22
Slik fordeler NMT trafikken seg pa en vanlig hverdag
Tele
Når alle bruker mobiltelefonen samtidig, kommer ingen fram

Har du noen gang sittet i bilko ut fra Oslo i firetiden en fredags ettermiddag? Det er håpløst. Framdriften kan måles i meter i timen i stedet for kilometer. Alt er bare kaos, og det ser ut som om alle skal samme vei som du.

Veiene ut fra Oslo har så absolutt et kapasitets problem. De er bygget for normal trafikk og greier ikke svelge unna altfor store mengder av gangen. Men kjører du utenom rushtiden, fungerer de utmerket.

Det er ikke bare på veiene det er trangt om plassen i Oslo.

Også andre steder har man fått kapasitetsproblemer. Det har sikkert du som bruker NMT merket. Til tider – f.eks. om ettermiddagen – kan det nesten være håpløst å komme fram.

På under tre år har vi fått hele 30 000 NMT-abonnenter i Sør-Norge, langt, langt flere enn vi regnet med da vi introduserte tjenesten. Bare i Osloområdet er det atskillig over 10 000.

Selvfølgelig er vi i Televerket glade for den uventet store suksessen, men den har også medført problemer. Vi har ganske enkelt fått for mange abonnenter til for få kanaler på for kort tid. I fjor ble det ført over 14 millioner samtaler på NMT. Det vil si innpå 40 000 samtaler hver dag. Rundt 35% av trafikken foregår innenfor en radius på 25 km fra Oslo sentrum. De fleste samtalene finner sted på formiddagen og mot slutten av kontortiden. Hvordan samtalene fordeler seg på en dag, kan du se pa figuren under.

Ekstra vanskelig blir det på grunn av de topografiske forholdene: Oslogryta er ikke akkurat som skapt for radiotelefoni – snarere tvert imot. Alle åsene rundt Oslo gjør nemlig at frekvensene vanskelig kan brukes pa nytt. Og det er ikke bare NMT abonnent[illegible] kniver om plassen. I luften over Oslo m[illegible] være rom for andre mobiltelefonsystem[illegible] fra radio, TV og lukkede kommunika[illegible]

Vi tredobler kapasiteten.

Selvsagt sitter vi ikke med hendene [illegible] om dette kom litt brått på oss. For tiden ar[illegible] hardt for å bedre situasjonen, og planene [illegible] I løpet av høsten øker vi kapasiteten me[illegible] neste år blir økningen like stor. I 1986 k[illegible] et helt nytt NMT system med tre gang[illegible] sitet.

Men inntil videre må du som bruker [illegible] være tålmodig. I denne overgangsfasen [illegible] å prøve å ringe mindre og fatte deg i kort[illegible] dags- og ettermiddagsrushet. Det er da [illegible] størst. Selvsagt er det lurt å utnytte den tide[illegible] sitter i bilko – det er bare det at når alle rin[illegible] samtidig, blir det vanskelig å komme fram. [illegible] prøver å begrense ringingen, blir det le[illegible] gjennom når det er noe viktig på gang.

Tele

Når alle bruker mobiltelefonen samtidig, kommer ingen fram

Figure 5 In 1984 the NMT-450 network was overloaded in Oslo, so the Norwegian Telecom had to make apologies for the situation by public advertisements

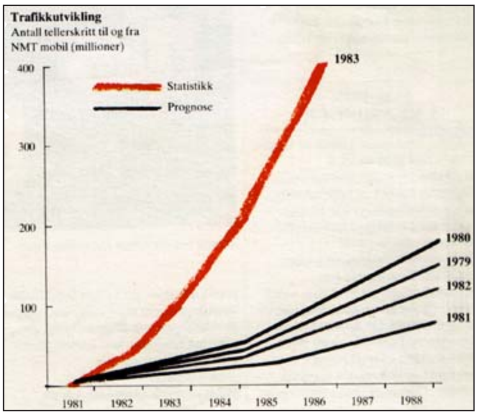


Figure 6 The real growth in the Norwegian NMT network was much higher than any prognosis

Also for the mobile switches and base stations there was close co-operation, mainly in the purchasing process, with similar consequences as for the mobile telephones. The price decreased and the functionality increased, particularly in those cases where there were competing suppliers.

The telecommunication operators of Denmark, Finland, Norway and Sweden launched their NMT networks in the 450 MHz band in 1981/1982. The functionality of the system was higher than for any other mobile telephone system in the world, and included handover between the base stations during the call as well as multi-country usage. In the autumn of 1982 the roaming function was available, which enabled a subscriber to use his mobile telephone when visiting another Nordic country.

The NMT was a success from the first day, and all the prognoses showed to be far too pessimistic. In Norway, NMT-450 was launched in 1981, and already in 1984 the network was overloaded in Oslo, as 35 % of all traffic was made within a radius of 25 kilometres from the centre of Oslo. Norwegian Telecom therefore had to temporarily suspend the subscription of new customers in the Oslo area.

The situation was similar in the other Nordic capitals. In 1983 the NMT Committee had started the development of a new system named NMT-900, which was to operate in the 900 MHz band where frequencies were reserved for public mobile use.

Experience gathered from the NMT-450 system, new requirements and available techniques, especially in the mobile telephone, were the basis for the development work which was finished in 1985. Better speech quality, hand portable terminals and higher capacity were the attraction for the users.

As of end 1986, NMT-450 had some 87,000 subscribers in Norway. NMT-900 was then launched as a complementary system, and the ban for new subscribers in the Oslo area could be suspended, provided they chose NMT-900. Since geographic coverage is one of the key success factors for any mobile telephone system, Norwegian Telecom in three years' time built NMT-900 to cover most of southern Norway. In June 1993 the subscribers in NMT-900 for the first time outnumbered those of NMT-450, each network then having 160,000 subscribers.

As a result of this development, the Nordic countries have always been in the top league regarding mobile penetration, with Norway and Sweden as the two leading ones. Internationally, the NMT project has been regarded as pioneering and has served as a model for mobile telephone services in more than 40 countries throughout the world.

The reasons for the success of NMT are several:

- The NMT system is based on well known and tested technology.
- The NMT system is robust and flexible.
- Properly specified interfaces make it easy to purchase equipment from several manufacturers.
- The mobile stations are not complicated. The manufacturers can use their

Figure 7 Eventually, mobile telephones decreased in size and were available in a variety of designs

resources to make the mobile telephones attractive to the users.

- Connection of new operators is very simple from a technical point of view.
- The NMT system is not patented, and all interested operators and manufacturers may adopt it.

However, the most important reason for the success is probably that the NMT system came at the right time and satisfied a basic need in the market.

GSM

CEPT (European Post and Telecommunication Conference) established already in 1982 a working group for developing a set of common standards for a pan-European mobile communication system. The standardisation work was carried further in ETSI (European Telecommunications Standards Institute) from 1988.

The main requirements of the system were to

- define the necessary interface specifications for an automatic pan-European network with ISDN interconnection, allowing full roaming capabilities in all participating networks
- provide basic telephone services with a quality at least as good as that of existing mobile communication systems, as well as a wide choice of non-voice services
- accommodate both vehicle mounted and hand-held mobile stations
- offer a traffic capacity higher than that in existing systems
- be able to co-exist with first generation mobile telephone systems, even at the same base station sites
- allow national variations in charging systems and rates
- lead to a total cost to the subscriber which is no higher than in existing systems.

In 1985 the basic principles of the system were agreed upon. The GSM working group concluded at an early stage that the new system should use digital technology. Just like NMT, the main function of the GSM switches would be to connect the mobile subscriber with other subscribers through the fixed network. One new feature, however, was that GSM was opened for several competing operators in each country, and that the subscriber's identity would be contained in a separate

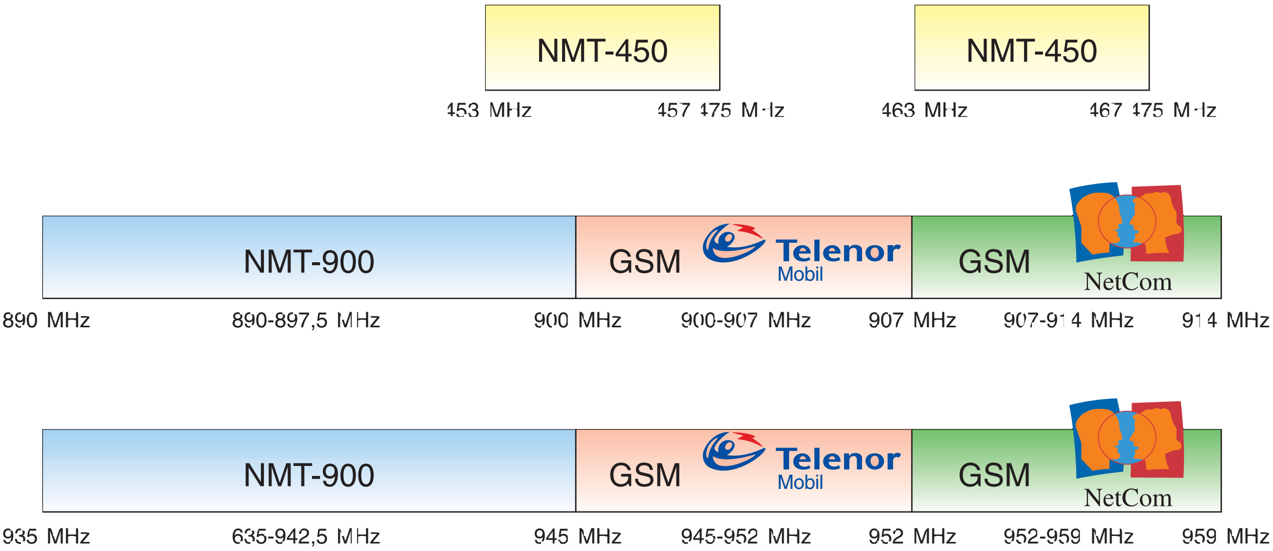


Figure 8 The frequency bands used for NMT and GSM

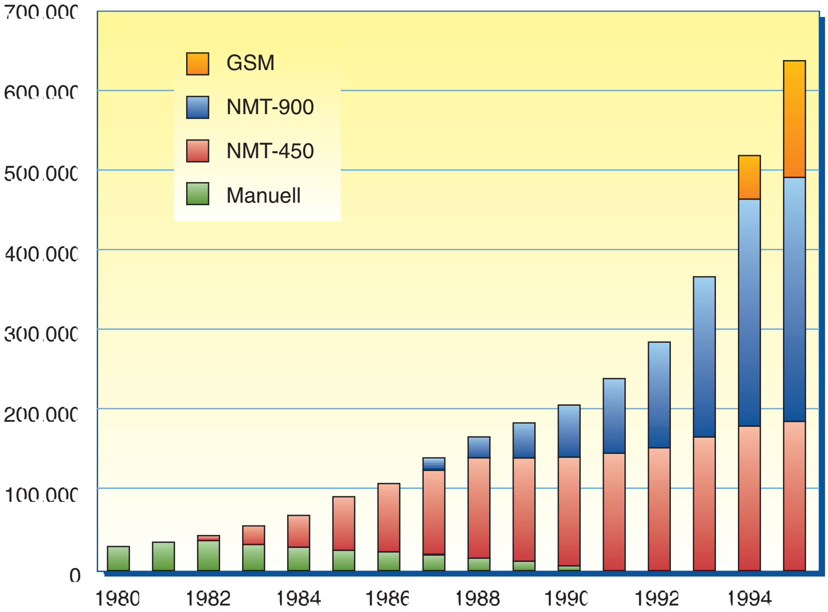


Figure 9 The number of subscribers in the mobile networks run by Telenor Mobile from 1980 until the first half of 1995

module (SIM), which is inserted into the mobile station.

The Nordic countries were (and still are) heavy contributors to both the technical GSM specifications and the administrative procedures and interfaces between the operators, based on long time experience and development of the NMT system.

After extensive development, in which several manufacturers took part, and testing of different candidate systems, ETSI agreed on the GSM standard. During the exhibition Telecom '91 in Geneva in 1991, GSM was for the first time in commercial operation.

After an extensive political process, the Norwegian Parliament decided in 1990 that two licences should be given to operate GSM networks in Norway, one of them to Norwegian Telecom Mobile and the other to a private operator. After an application process the licences were given in December 1991 for a period of 12 years.

GSM was put into commercial operation in 1993, with two competing operators: Telenor Mobil AS (which also operates the two NMT networks) and the private operator NetCom GSM AS. This digital system opened up for new service opportunities for the customer, such as data and short message communication as well as roaming capabilities with a large number of countries around the world. By the end of 1994 the two mobile operators had altogether nearly 130,000 subscribers, which it took 6 years to achieve in NMT.

GSM has been adopted throughout Europe, and also by a number of countries in other parts of the world. There are at present (December 1995) around 165 GSM network operators in 88 countries. More than 90 GSM networks are in operation and the number is steadily increasing.

Paging services

Norwegian Telecom introduced radio paging as a public service in 1984. Two alternative systems had been considered, one with separate transmitters, the other utilising the FM broadcasting transmitters. The system chosen was one with separate transmitters, based on the POCSAG code recommended by CCIR.

The first system was a numeric paging system which was only capable of transmitting alerting calls and, as an option, numeric information. The subscribers were typically requested to make reply calls to telephone numbers disclosed on the pager. The paging service soon became popular and Norway achieved one of the highest percentage of paging penetration in Europe.

An alphanumeric paging service was launched in 1991. An alphanumeric message may consist of up to 128 characters. Transmission of news messages and stock exchange information are examples of value added products which can be sent through the alphanumeric paging network.

Paging services were liberalised in 1989, but no competitors have so far challenged Telenor.

Mobile data

From 1990, Telenor is operating a Mobitex service in the more densely populated areas of Norway. This is a public packet switched mobile data service, which is particularly well suited for companies and public utilities having much of their operations in the field. Mobitex is offered in addition to the line switched GSM data service. The mobile data services can be adapted to the customer's requirement by means of suitable application software.

Mobile satellite communication

During this century, the Norwegian shipping business has maintained its position as one of the leaders in the world. This market enabled Norwegian Telecom to develop an efficient maritime radio communication service. Direct long distance traffic with ships was made possible by means of HF radio communication as mentioned earlier, but the transmission quality can be very variable. In addition, the traffic growth resulted in congestion in the maritime high frequency bands.

After extensive preparatory work, the intergovernmental organisation INMARSAT was formed in 1979 to

Figure 10 Paging represents a convenient and lightweight way of being available

establish and operate a world-wide maritime satellite system. Norway was very active in the preparation of INMARSAT. Norwegian Telecom was one of the pioneers of maritime satellite communication, and established, in co-operation with the other Nordic telecommunications administrations, the first maritime satellite earth station in Europe. The station opened in 1982 at Eik, not far from Stavanger, shortly after INMARSAT had become operative. Based on an agreement with British Telecom and Telecom Singapore, automatic telephone and telex services could be offered to the shipping industry at unified prices all over the world.

Maritime satellite communication has proved to be very successful and profitable. The number of services has increased and includes aeronautical and land mobile use. Telenor continues to be one of the largest shareholders of INMARSAT.

Towards a personal communication service

The advances in mobile communications have been enormous and the development continues at an even greater pace. It is estimated that the mobile telephony service has a potential of 80 per cent penetration compared with about 60 per cent for the ordinary telephone service, and can then truly be named personal communication. In Norway this evolution is well under way, stimulated by lower terminal prices, attractive services and competition. The penetration in Norway is today (December 1995) about 23 per cent.

GSM will gradually take over as the dominant mobile telephone system in Norway, but the demand for NMT will exist for many years because of the good geographical coverage of the system. Because the frequency resources for GSM in the 900 MHz band are limited, there will soon be a need to start utilising frequencies in the 1800 MHz band (DCS 1800).

As regards satellite communication, several low earth orbit satellite systems are planned for communication via portable terminals or mobile telephones. In this way it would be possible to cover areas which will not be profitable to cover by terrestrial systems. Also, satellite systems may be combined with terrestrial mobile systems, thus making it possible for the subscriber to change automatically between the systems.

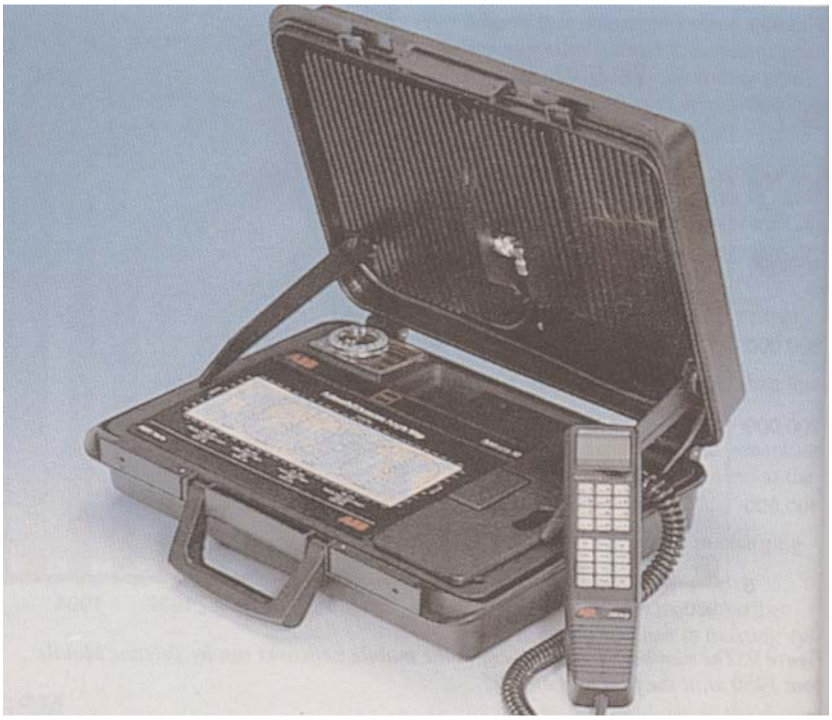


Figure 11 Mobile satellite telephony provides global access were land mobile systems are not available. Here is shown a "lightweight" Inmarsat M phone from the Norwegian company NERA

Abbreviations

MF:	Medium Frequency
HF:	High Frequency
VHF:	Very High Frequency
UHF:	Ultra High Frequency
NMT:	Nordic Mobile Telephone
GSM:	Global System for Mobile Communication
ISDN:	Integrated Services Digital Network
INMARSAT:	International Maritime Satellite Organisation
FM:	Frequency Modulation
CCIR:	International Radio Consultative Committee

DECT field trial in Førde

BY BJØRN ERIK ESKEDAL, JOAR LØVSLETTEN AND JAN TORE DEILKÅS

1 Background

Access via radio provides the operator with an attractive alternative to using conventional copper loops in building up own infrastructure. Use of radio can provide new customer services like mobility and is cost effective compared to wired solutions in several environments. In order to gain experience with radio access technologies and evaluate subscribers' interest in local mobility Telenor has established a DECT trial network in the small community of Førde. Førde is located in a mountainous area in Western Norway and has approximately 9,000 inhabitants. The main reason for choosing Førde for the field trial was that residential, down-town, business and industrial areas were all concentrated within a small local area giving the possibility of covering the area with a reasonable number of base stations and thereby allow roaming between different types of environments. The possibility of using the same handset at home, at the office or in public areas represents an important step towards achieving personal communication services. Seen from the market point of view this trial gives Telenor a special opportunity for studying the customers' reaction to this new service. Seen from a technical point of view the trial covering indoor and outdoor as well as different types of environments gives valuable information on how well the DECT technology can function in a multi operating environment.

The main reason for choosing a DECT system was that it is a European standard with a dedicated frequency band allocated for the system (1880 – 1900 MHz). It supports a standardised radio interface capable of providing interoperability between equipment from different manufacturers. It has the capability of supporting local terminal mobility and is developed to connect to target networks like PSTN, ISDN, X25, IEEE 802, LAN, GSM and support a wide range of voice and data services.

The DECT network has been in operation since mid-1994, but the trial customers were not connected to the network until December 1994. Market tests are to be conducted throughout 1995. At this stage (May 1995), main focus has been on providing terminal mobility within the community. However, tests of using DECT technology as a fixed radio access solution replacing copper the last drop is also of interest and is expected to be conducted later.

2 Network configuration

2.1 System architecture

The DECT trial system used in Førde consists of 160 base stations controlled by a radio exchange. 240 pocket sized handsets are connected to the system. The base stations are connected to the radio-exchange using two (unshielded) twisted pairs, each with a capacity of carrying four 32 kb/s ADPCM channels. The communication is a proprietary half duplex ISDN 2B+D transmission. The radio exchange is connected to the public local exchange through 8 multiplexer units, each handling 30 speech channels. The multiplexer units are connected to the radio exchange with a 2 Mb/s connection.

In order to integrate the business users with their local PABX numbers into the DECT system, their fixed telephone lines were connected directly into the multiplexer units.

All base stations are remotely powered from the radio exchange, using the same copper cable pairs as used for data transmission. In order to feed enough power to the most remote base stations, it was necessary to increase the voltage up to 96 V. As a general rule measurements performed indicated that base stations located more than 1 km from the radio exchange required more than 48 V. In order to supply the 96 V, an external rectifier of 550 W were connected in series with the central secured power source.

Figure 1 shows the configuration as described above.

3 Services offered

Approximately 50 households and 55 business/industry companies are participating in the trial. The number of handsets used in a family varies typically between 1 and 3. By allowing in some cases several family members to have their own handset with a dedicated telephone number a personal service is offered to the customers. The number of handsets offered to a company varies between 1 and 11. Calls can be made and received within the entire covered area (approx. 2.2 sq. km) and seamless handover is performed between all 160 base stations. Some trial customers both live and work within the coverage area, benefitting from the possibility to use the same handset both at work and at home. During the test period private customers only pay the same tariff as fixed telephone tax. The same offer is given to business users. In addition, all normal PBX functions are provided allowing business users for instance to make free internal calls within the covered area.

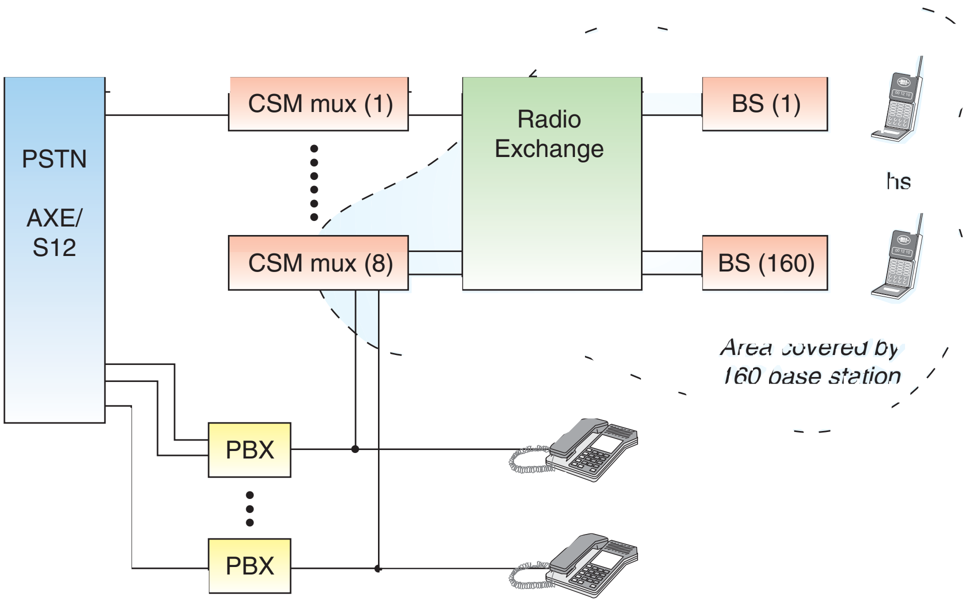


Figure 1 Network architecture

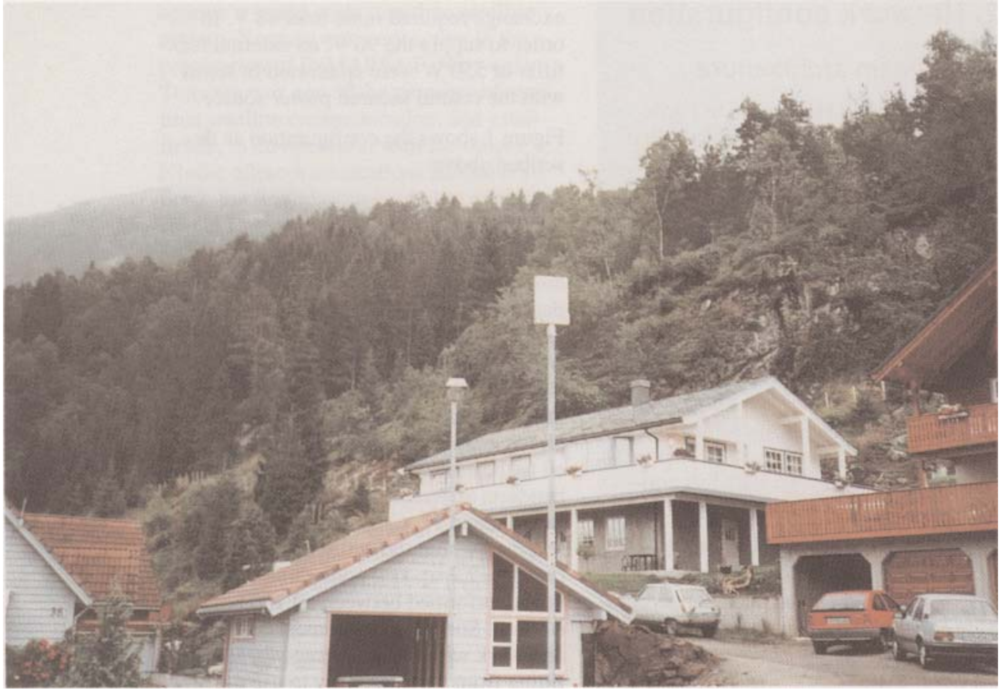


Figure 2 Typical base station installation in a residential environment

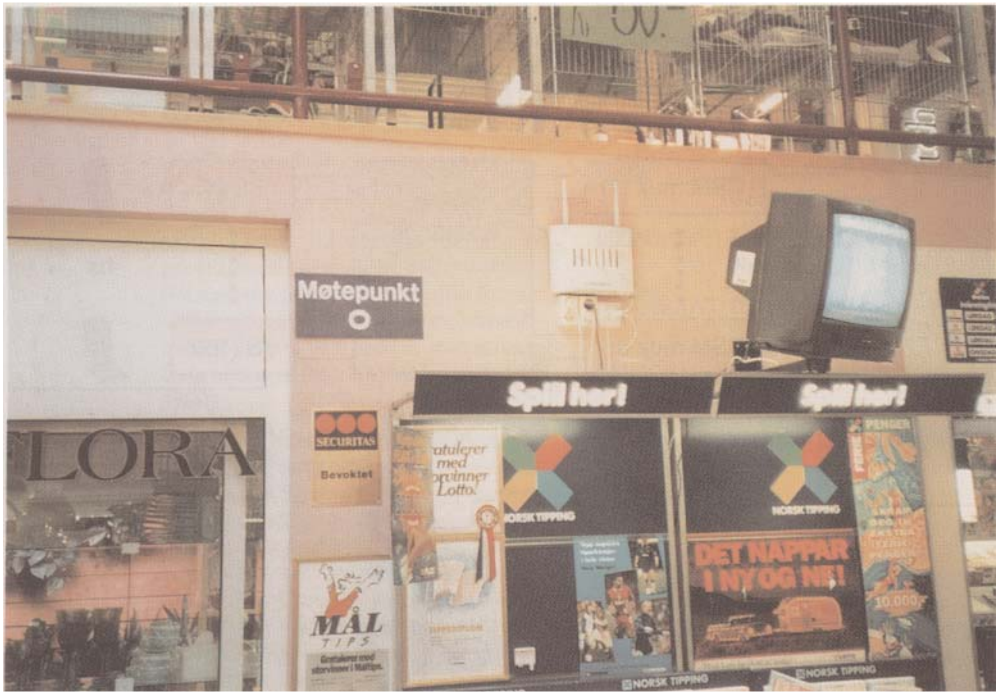


Figure 3 Base station site inside a shopping centre

4 Radio planning strategy

The idea behind the trial was to build up a network of base stations giving full radio coverage within a certain area including residential, business, industrial and shopping areas.

Residential area

The selected area is approximately 0.9 sq. km, hilly, with some vegetation and consists of approximately 700 residences including detached houses with own gardens as well as apartments. Within this area both outdoor and indoor coverage is provided. All base stations are mounted outdoor using existing lamp poles, walls of buildings or in some cases by deploying new 3 – 5 m masts (see Figure 2). To minimise digging and labour costs, the base stations have been connected as close as possible to end distribution points attempting at the same time to get a best possible site location which on one side has a minimum impact on the environment and still is a good point seen from a radio propagation point of view. Due to the low radiated output power (24 dBm, 250 mW) and poor receiver sensitivity level requirement (–83 dBm) for both the base station and handset the cell range in an outdoor environment is typically limited to a few hundred metres. To provide indoor coverage from base stations located outside as done in Førde an additional penetration loss through the building must be calculated, limiting the range to typically less than 100 metres. Different types of directive antennas have therefore been used in parts of the residential area to increase the cell radius from typically 40 – 60 m to 200 – 400 m. A total of 64 base stations have been installed covering the residential area.

Town area

This area is approximately 0.4 sq. km and consists of public buildings, shops and offices. Within this area complete outdoor coverage is provided by base stations sited on walls and roofs of existing 3–5 storey buildings or other existing structures like telephone boxes. In addition, all public buildings like post offices, banks, department stores, etc. are covered as well as all trial business customers living in this area. Both indoor and outdoor base stations are used for providing continuous coverage. A typical indoor cell site installation covering part of a shopping centre is shown in Figure 3.

Table 1 Environment characteristics and system planning aspects of the DECT field trial performed at Førde

	Residential area	Downtown area	Business/industrial area outside the town centre	Recreation area
Area characteristics	- 0.9 km^2 - 700 residences - both hilly and flat areas with some vegetation	- 0.4 km^2 - shops, public buildings (3–6 storeys) - flat area	- 0.4 km^2 - massive buildings with thick concrete walls - flat area	- 0.5 km^2 - fields, open green areas, sporting area
Planned coverage	- outdoor - indoor in most rooms except basements	- outdoor - indoor in public areas - indoor for all business customers - partly indoor coverage in shops	- outdoor - indoor in areas of importance for the trial customers	- outdoor - indoor in public areas
Installed base stations	- 64, all outdoor	- 15 outside - 40 inside	- 11 outside - 19 inside	- 3 outside - 6 inside
Subscribers	- 52 families	- 27 business customers	- 16 business customers	- 4 business customers
Number of handsets	- 90 handsets	- 74 handsets	- 58 handsets	- 10 handsets

Business/industrial area located outside the town centre

The business/industrial area located within the coverage area but outside the town centre is approximately 0.4 sq. km and consists typically of a lot of massive constructions with large open halls and a lot of reflecting objects. Almost blanket outdoor coverage across the area is provided. However, due to the very limited range inside some difficult areas, e.g. in cellars, blank spots with no or poor coverage exist. Both indoor and outdoor base stations have been used to give the wanted coverage.

Recreation and open areas

The area belonging to this category consists primarily of recreation and open areas, a culture house, a sporting centre and schools. In total, the area is approximately 0.5 km2. Blanket outdoor and indoor coverage in public areas is provided.

Table 1 summarises some environment characteristics and system planning aspects of the DECT field trial.

5 Experiences from the ongoing trial

Most existing DECT products have been designed and targeted for indoor applications. This is also the case with the current Ericsson version implemented in Førde. However, keeping in mind the potential a system like DECT may have in an outdoor environment both used as a technology providing local mobility and used as a fixed radio access solution replacing copper the last drop, one of the main technical goals was to get a better understanding of the strengths / weaknesses and improvements needed on current versions of the DECT system in order to operate in both indoor and outdoor environments.

Indoor

When planning indoor coverage, several aspects must be taken into consideration in order to minimise the number of base stations and still achieve high speech quality. This includes accurate positioning of the base stations, 3-dimensional planning strategy, knowledge of the interior constructions and types of reflecting objects and materials of ceiling, walls and floors attenuating the signal strength a lot. Experience from the trial has given valuable information about the expected cell radius in a variety of different indoor environments ranging from more than 50 metres in open hall areas to less than 15 metres in heavily reinforced areas. As a result of the limited cell radius careful site planning can reduce the number of indoor base stations by more than 30 %.

Outdoor

A lot of subjective tests have been performed in different types of environments giving valuable information of the system performance. Some positive experiences with the system are:

- The speech quality is in general good (e.g. better than GSM) as long as there is a free line of sight path between the base station and the terminal and no major reflecting objects are in close vicinity.
- Good speech quality has been obtained more than 400 m from nearest base station (without using directive antennas) and approximately 1 km from base stations using directive antennas.
- Even though the system is not developed for use in a car, tests show that it is no problem communicating and performing handovers while driving in a car. However, a certain degradation in quality is experienced.

Some negative experiences are:

- A reduced speech quality has been observed in open square areas typically surrounded by a lot of reflecting buildings even when there exists a free line of sight path between the base station and the terminal and high average received signal strength levels are measured. Multipath measurements done in similar square areas [2] indicate long delay spreads which may explain the reduced performance.

- The link quality is highly dependent on whether line of sight is available or not.
- The link quality is dependent on whether the user is moving or standing still, particularly if no line of sight exists between the base station and the handset.
- User's orientation and positioning of handset relative to the base station affects the link performance.

7 Conclusions

A DECT field trial has been established in the small community of Førde to gain experience with the DECT technology, evaluate its potential for providing local mobility across different types of environments and survey customers reactions and interest in local mobility.

The experience so far in the Førde trial shows that the DECT technology is a strong candidate for providing speech services and mobility in indoor domestic/business/industrial environments. However, for providing outdoor local mobility the technology is still immature and too sensitive to radio propagation conditions. Improvements in range e.g. achieved by use of repeaters, use of advanced diversity techniques or channel equalising to combat multipath situations seem necessary [2]. On the other hand, with an improved air interface the high capacity of DECT, the variety of services supported by the standard and its simple/flexible network structure makes it a very interesting candidate both used as a pure copper replacement alternative connecting subscribers to the fixed network and used as a public radio access solution providing local mobility. Both manufacturers and standardisation groups are continuously working on improving the equipment and developing new packages to meet the future customer demands for personal communications. A particular interesting solution is the development of dual mode DECT/GSM, DECT/DCS-1800 handsets. This enables the user to combine the strengths of GSM and DECT into one handset and for instance use DECT as the local radio delivery system with low tariffing and extended set of services and switch to GSM modus when wider coverage is required.

8 References

1 Eskedal, B E. Examining the performance of DECT/Ericsson in a Multi operating environment. In: *IIR conference, developing and exploiting wireless local,* May 1995.

2 Rækken, R H, Eskedal, B E. DECT performance in multipath environments. In: *Nordic radio symposium (NRS 95),* Saltsjöbaden, April 1995.

The DECT system

BY BJØRN ERIK ESKEDAL

1 History

The reason for developing a common European standard for cordless communication was motivated by the complicated situation on the cordless market in the 1980s.

A lot of incompatible cordless telephone systems existed and were in use in Europe. Many products were inexpensive telephones illegally imported from the far east. In addition, some European products as the analogue CT1 system and a digital FDMA system (CT2) developed in the UK, were adopted and used in some countries. However, all these systems were relatively simple, offering only some of the facilities and services a user would require from a complete cordless system. Based on this situation it was decided to develop a standard cordless radio interface capable of providing interoperability between equipment from different manufacturers and meeting the growing demands and expectations for cordless voice and data communication. A single frequency band was dedicated for this standard called Digital European Cordless Telecommunications (DECT) in the range 1880 – 1900 MHz. In 1988 ETSI established a committee to work on the standard. In June 1991 the specification was completed and published a few months later. The DECT standard got its final approval as ETSI standard in June 1992.

2 General description of the DECT system

2.1 Network architecture

The DECT system has been designed primarily to provide low power two-way tetherless voice and data communication between small inexpensive portable parts and fixed points (radio base stations) for distances up to typically a few hundred metres, supporting only low user mobility. The system serves only as an access network specified to connect to target networks such as PSTN, ISDN, X25, IEEE 802, LAN and GSM.

The interfaces to the host networks are not directly a part of the standards but are explicitly included in the standards to ease implementation. It is physically possible to connect DECT to these networks in a number of ways and at different distribution points. It is further possible to physically integrate the DECT specific functionality and the functionality of the target networks into common building blocks (e.g. WPBX solution) or to keep them separated. Figure 1 shows one possible physical implementation of the DECT system including the three basic elements of the system.

Central Control Fixed Part (CCFP)

The central control fixed part has the overall control of the radio resources and radio connections in the DECT system. In order to control the movement of users between radio base stations (RBSs), mobility management and call related functions are taken care of by this unit. Also interworking functionality is required to convert protocols from the format used in DECT to the format used in the attached networks. All incoming and outgoing calls are routed via the central unit to the RBSs.

Radio Base Station (RBS)

A DECT system may have several hundred radio base stations connected to the CCFP, each transmitting with a peak power of 250 mW giving radio coverage across a limited area. All functions required for supporting the physical transmission across the radio path is located in this entity. A standard DECT RBS has the capacity to handle 12 simultaneous calls. On average, it offers 5 Erlang speech traffic (32 kb ADPCM) with a blocking rate less than 0.5 %. In DECT all RBSs share the same resources. That means all 10 carriers allocated for the system can be used by all RBSs. The coverage area of an RBS varies a lot depending on the characteristics of the environment. However, in general, an RBS will provide indoor coverage ranging from 20 – 60 metres from the transmitter and up to 500 metres outdoor if implemented with a standard omnidirectional antenna. If a directive antenna is used, the outdoor range may be increased to several kilometres. In contrast to GSM, DECT uses no equalisation or error correction protocol for speech transmission making it vulnerable to fading dips and multipath effects. To combat the worst effects, however, each RBS is typically equipped with two antennas transmitting and receiving on the best one. The connection between the RBSs and the CCFP can be two or three twisted pairs of wire each carrying 144 kb/s (S) providing four 32 kb/s ADPCM simultaneous half duplex calls. Alternatively, a 2.048 Mb/s connection supporting up to 30 simultaneous calls is possible. Depending on the diameter of the cable and powering used, the distance from the CCFP to the RBSs may be up to 3500 m.

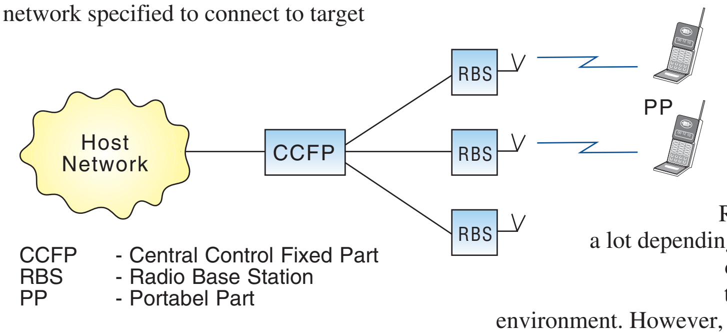


Figure 1 Typical DECT network architecture

Portable Part (PP)

The DECT portable part contains all functionality needed to support the user with similar teleservices as provided by the attached target network everywhere in the coverage area. During a call the user is free to move around throughout the covered area without losing the connection. All radio transmission functions needed to communicate across the standardised air interface towards the radio base stations are included in the terminal. In addition, a DECT portable part has been provided with a lot of intelligence. Calls and handovers are for example controlled in the terminal. This decentralised concept makes it easy for the system to quickly react to changes in the radio environment.

2.2 Multiplexing technique

The DECT system uses a Multi Carrier-Time Division Multiple Access / Time Division Duplex (MC-TDMA/TDD) scheme to transmit the information. Ten carriers are allocated to the system in the frequency band 1880 – 1900 MHz with a carrier spacing of 1728 kHz. The first 12 timeslots are used for transmitting in the direction from the base stations to the handsets. The following twelve timeslots are used for transmitting in the opposite direction, giving twelve duplex physical channels for each carrier. Each frame consisting of 24 timeslots lasts for 10 ms giving a throughput of 100 frames per second (see Figure 2a). An RBS may operate on all 120 duplex channels in the

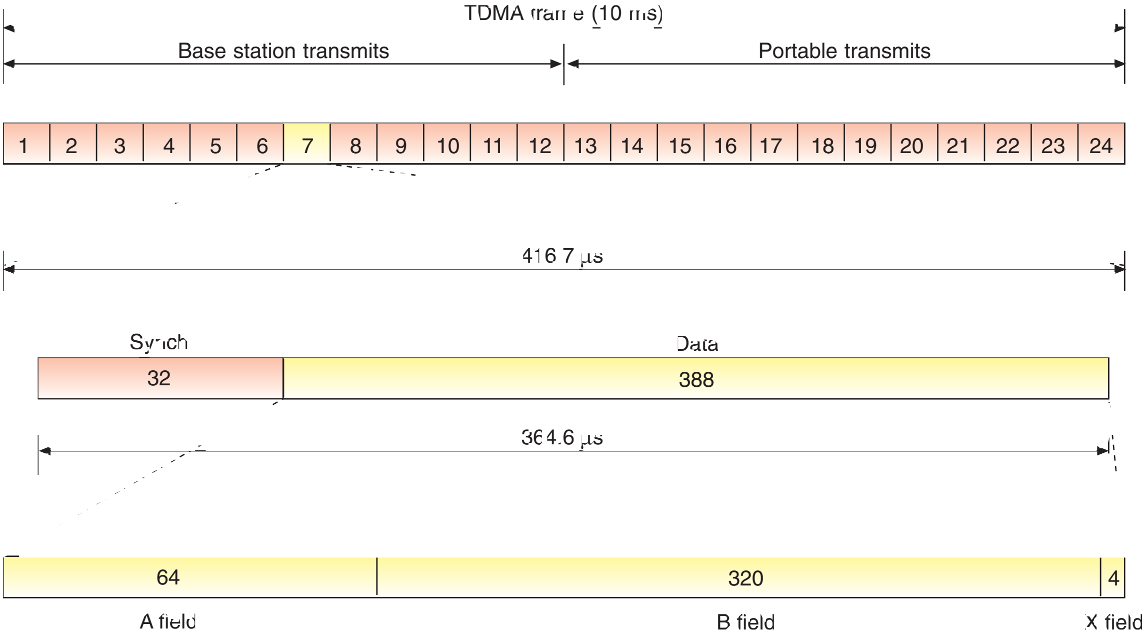


Figure 2a TDMA frame structure

system and is capable of switching carrier from one timeslot to another.

In DECT control and data information are multiplexed together to one physical packet which is transmitted over the air interface. At the beginning of a packet a 32 bit synchronisation field (S field) is sent followed by a 64 bit A-field, containing control/signalling information providing a net channel signalling rate of 6.4 kb/s. The B-field carries user data (320 bits) during a call, providing 32 kb/s of user data. The timeslot structure for DECT is shown in Figure 2b.

With this TDMA frame structure, DECT has the options of offering simplex and duplex communication as well as providing symmetric and asymmetric links. The simplest duplex service uses a single pair of evenly spaced timeslots to provide a 32 kbit/s digital information channel capable of supporting one speech conversation or a similar low rate digital service. Higher data rates are achieved by using more timeslots, and lower rate services are provided by using half slot data bursts.

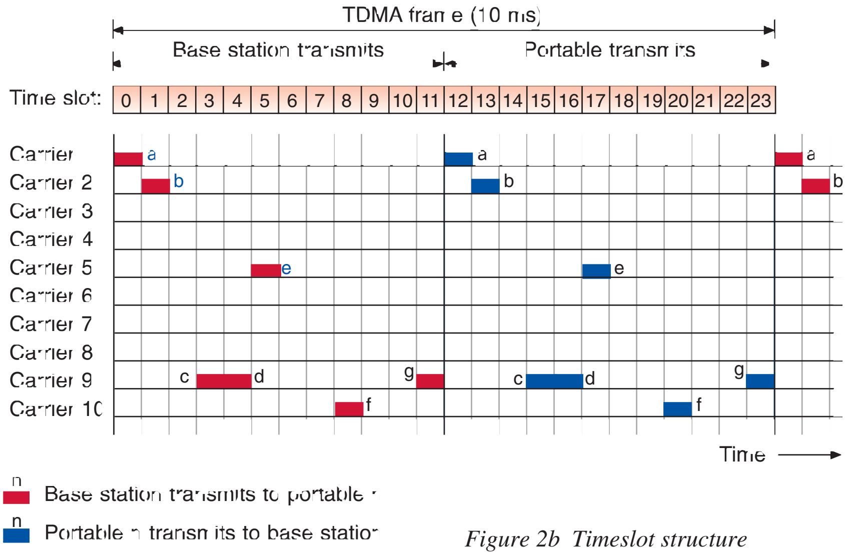


Figure 2b Timeslot structure

2.3 Dynamic channel selection

DECT is developed to provide an overlapping radio coverage area using a lot of low power RBSs, each covering a small area. By limiting the cell radius to a few hundred metres a shorter reuse distance is achieved and thereby an increase the system's total traffic capacity. However, seen from a radio planning point of view, due to large local traffic variations it is very difficult to plan effectively the expected traffic load on each RBS and predict the number of channels needed. A fixed channel allocation scheme as used in mobile systems would in general require dimensioning each RBS after a worst case model to cope with peak situations. The result is poor utilisation of the scarce radio spectrum available. Taking this into account when designing the DECT system a dynamic channel selection procedure has been implemented making it possible in principle for all handsets and RBSs to communicate on all channels in the system. For each call the portable part selects the most suitable channel at that time for the connection. That is a channel with a high signal/noise ratio. If, however, the signal level gets degraded or the channel is stolen from another handset, a new available channel is chosen immediately using a seamless handover procedure as explained below, preventing the user from noticing any impaired speech quality.

The dynamic nature of allocating resources after demand makes it simple to extend the system by simply adding new RBSs and let the system do the rest. In addition, the DCS procedure makes it much simpler for different uncoordinated DECT systems and system operators to coexist utilising the same radio resources.

2.4 Signalling and control information

Frequent system information and access rights identification is transmitted from every RBS making it possible for the handsets by just listening to the information to identify any system or RBS within reach and lock to it. The system information includes synchronisation information, capabilities of the system and radio carriers in use. The access rights information lists the available services from different operators accessible from the system within reach. If the handset has identified a valid system, it locks to the strongest RBS. In this idle state it listens every 160 ms for a possible paging call from the system (in the signalling/control field of a timeslot). If a call is initiated by the portable part or it has been requested by a paging call, the dynamic channel selection procedure finds a free physical channel and transmits a message on the best one. By listening for a paging message every 160 ms the call delay does not exceed 100 ms on average.

2.5 Handover procedure

Since the coverage provided by one RBS in DECT is limited, a mobile user will often move outside the coverage area of one RBS entering an area covered by another RBS. In this case the signal level from the serving RBS gets too low, requiring the handset to establish a new and better channel to the next RBS. This feature of changing the physical channel from one RBS to another is called intercell handover. In other situations co-channel interference from other DECT users may cause bit error increases. In this case a switch of the call in progress from the interfered channel to another less interfered channel of the same RBS is required. This feature is called intracell handover. Handover (both intracell and intercell) is used in DECT to maintain a quality level comparable to what the fixed networks can offer. Whenever the quality level gets degraded the portable part sets up a new link in parallel to the old one on another and better timeslot. When the new link is established the new RBS (in case of an intercell handover) informs the CCFP to make a seamless switching from the old to the new radio link. In this way the user does not notice when a handover takes place. The decision to perform handover is taken by the portable part and is based on radio link quality measurements as described below.

2.6 Measurement of the radio link quality

In DECT several indicators to monitor reception quality may be implemented in the portable or the fixed part. The two most used mechanisms for this are signal strength measurements and bit error control.

Signal strength measurements

During a call the portable part measures the signal strength from the serving RBS and neighbouring RBSs. These measurements are recorded in an ordered list of physical channels with the greatest signal strength level. The channel list is updated at least every 30 seconds. In this way the handset is prepared to make a quick change of channel to another RBS if the signal level from the serving RBS gets weak. In addition, the portable part scans all channels finding the ones with the lowest received signal strength level. This list of “free” channels is used to find an appropriate channel for call set-up or to perform a handover to.

The A field CRC

To provide error control on all A-field data (control data), 16 CRC bits are mapped to the A field (see Figure 1). In this way the receiving part, handset or RBS can supervise the link quality in both directions. The handset is in charge of making the decisions on what to do if the quality gets bad (handover decision). This means that if the RBS detects A-field check sum errors this is reported back to the handset which in turn decides what to do.

2.7 Security mechanisms

In earlier generations of analogue cordless systems the security mechanisms were often poor making it possible for unauthorised persons to break into the system and listen to private calls or communicate via private RBSs and thereby make “free calls” on the expense of the owner of the RBS. As a result and to counter for these and other threats, strong security features for the DECT system have been implemented.

The security mechanisms for DECT have two aspects. One is to provide protection of the transferred information across the radio link (encryption). The second aspect relates to securing that only authorised persons or terminals may connect to the network (authentication).

Encryption

To ensure privacy for user information (speech or data) and signalling information over the air interface between the handset and the fixed part, a standardised encryption algorithm has been specified. Both the handset and the fixed part establish a common cipher key. During a conversation or data transaction this key is used to encrypt the data by generating a stream of scrambled bits. The process is reversed at the other end to regenerate the data. Encryption of user data is not a mandatory function in the DECT system but is envisaged.

Authentication

Authentication of a handset:
All handsets in a DECT system have a unique identity which is checked by an authentication process. This is to prevent stolen or non type approved handsets to be used in the system and to prevent one DECT subscriber from using another subscriber’s identity to avoid call charges. If the telephone gets lost or stolen, the associated identification number (programmed in the handset) may simply be removed from the system making unauthorised calls impossible.

Authentication of a handset is usually done during set-up. A random number is sent across the air interface to the handset. Both the handset and the controller unit encrypt this value with an authentication key. The handset then returns the encryption result to the fixed part which in turn compares the results. If the compared results are identical the handset is accepted. If not, the call is cleared.

User authentication:
If a user authentication service is required in the DECT system the user has to manually enter a personal identity into the handset at the start of a call. This service allows the system to check a personal code associated with the user ensuring that only the right owner of the handset knowing the code of the handset gets access to the system.

Authentication of the fixed part:
Fixed part authentication provides a means to the user to check the system identity preventing other DECT systems operating in the same area to load infor-

mation into the handsets (such as a Zap code) which may render the portable unstable.

3 Applications

DECT is an access technology with a specified common air interface, able to connect to a lot of existing networks. This means that the DECT system must support most services and facilities offered by the target networks. The system has been designed to operate in several areas of applications ranging from simple single-cell domestic cordless use to large complex wireless PBX systems. It is particularly targeted at the following applications:

- domestic cordless use
- wireless business (PBX) use
- wireless local loop
- wireless LAN access
- public (telepoint) access
- evolutionary applications.

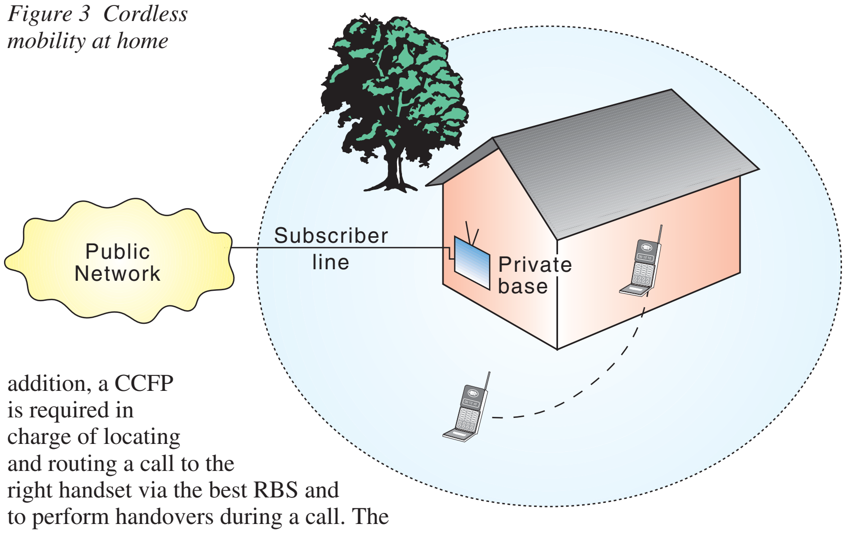
3.1 Domestic

This is the simplest application of the DECT system and can be seen as a radio link extension of the fixed telephone line. In this configuration DECT has the same basic functionality as earlier generations of cordless telephone systems making it possible for the subscriber to make and receive calls within the home and close vicinity. Both the handset and the RBS are typically owned by the private subscriber, and the DECT fixed part is simply connected to the subscriber's network interface as a plain old telephone. Since this is a very simple application with only one base station the interfacing functions (located in the CCFP in larger systems) and functions of the RBS are combined and co-located. Enhanced features (optional in the DECT specification for domestic use) allowing more than one handset to connect to the RBS and allow internal communication via the RBS are likely to be included. Figure 3 shows the domestic application.

3.2 Wireless business (PBX)

Figure 4 shows a possible configuration of DECT used for business cordless communication. Compared to the domestic application this application is far more complex requiring much more functionality of the DECT system. To cover the complete area of a large business system a lot of overlapping RBSs are needed. In addition, a CCFP is required in charge of locating and routing a call to the right handset via the best RBS and to perform handovers during a call. The CCFP may be integrated as a logical part of the PBX or connected to it as a separate unit. Seen from the users point of view communication and movement across the complete coverage area is achieved making a seamless handover to the best RBS when required. Normal PBX functions will be supported, giving the user similar services as offered by a wired PBX attachment.

Figure 3 Cordless mobility at home



3.3 Public (telepoint) access

DECT products on the market offering local mobility are primarily designed and targeted for indoor use. A general weakness with these products are their sensitivity to radio propagation conditions. Improvements in range, e.g. achieved by use of repeaters, use of advanced diversity techniques or channel equalising to combat the difficulties experienced are however looked into by several manufacturers and are expected to be available in 1996–1997. If DECT is used for this type of application the operator will typically install a lot of RBSs giving radio coverage at public places, city centres, hot spot areas, etc. As a basic feature both outgoing and incoming calls from a handset should be supported. Also the system should accept handsets which are manufacturer independent and allow the possibility of using the same handset both as a private cordless phone at home and as a public terminal when moving outside the home environment.

3.4 Wireless local loop applications

The wireless local loop applications involve using the DECT air interface to replace the wired link to the subscribers. Applications replacing the last few kilometres of cable by a radio link may be a cost effective solution for a network operator and a quick and efficient way of connecting new subscribers since no additional infrastructure is required between the subscriber and the operator's network. By using directive antennas either at

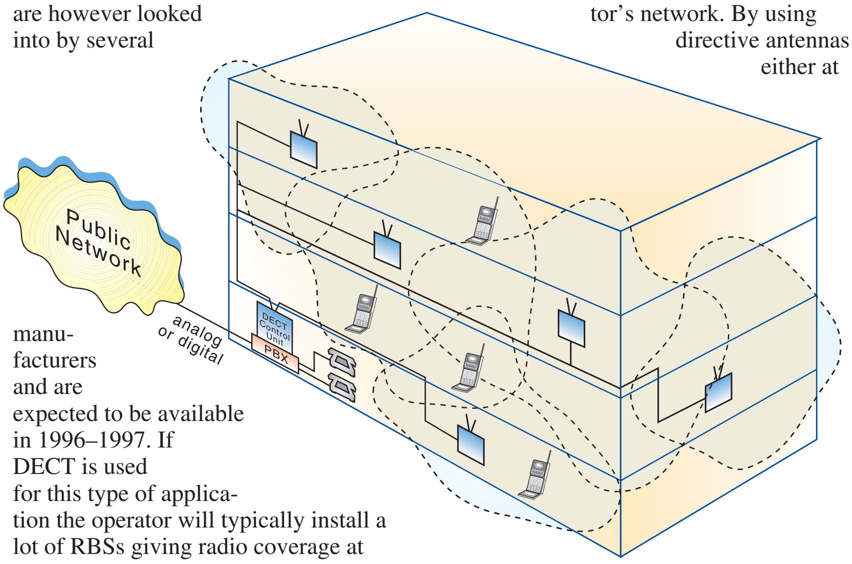


Figure 4 Wireless PBX application

the RBS side, at the terminal side or both, and in addition have the option of using repeaters the radio range may be increased to several kilometres. At the subscriber's end two solutions are envisaged. One is to install a fixed access unit (FAU) on the wall or roof of a house with a standard subscriber interface to connect equipment such as telephones, modems, PCs, etc. Seen from the subscriber point of view this solution offers no mobility. The second alternative is to install a fixed radio unit (FRU) with the possibility of working both as an FAU and in addition function as a repeater providing indoor coverage inside the home and in public environments (see Figure 5).

3.5 Wireless LAN access

One target network type with which DECT shall interwork efficiently is local area networks based on Ethernet and Token Ring. Packet switched bearer services, capability to support asymmetric or unidirectional data enables effective data transfer. Potential applications include fixed and portable station attachments to a LAN using the DECT air interface. In addition, the DECT network can be configured as a cordless gateway between two LANs. In Figure 6 a possible DECT wireless LAN application is shown. The network consists of terminals with a DECT portable radio termination part communicating with a wired LAN through RBSs. The HUB operates as the controller unit with control of the RBSs. Using the time division multiple access scheme in DECT to establish a connection oriented link between the server and any client workstation the net throughput is up to 552 kbit/s at a bit error rate not exceeding 10^{-3}. Higher capacity can be achieved by adding several base stations within the same cell.

3.6 Evolutionary applications

In the long term there will be a growing demand for full personal communications services. This means the possibility for a person to communicate whenever he wants at any place from a light handheld pocket sized terminal. A business user will typically require to use the same handset for incoming and outgoing calls when working in other offices of the same company or while attending meetings at hotels, etc. In addition, future DECT systems should support mobility across several environments making it possible for the user to use his handset at home, work, and in public areas as well as allowing roaming between various network operators and provide incoming calls everywhere as a basic feature. Combined dual mode DECT/GSM handsets is an additional very interesting future service feature already being looked into by major manufacturers of mobile and cordless equipment.

A combined DECT/GSM solution would enable a user to communicate through a private or public DECT network in local environments, e.g. indoor at home or at the office and in local areas, and switch to the GSM network when leaving the DECT coverage area. In this way the user would benefit maximally from the strengths of each system concept and only need one terminal.

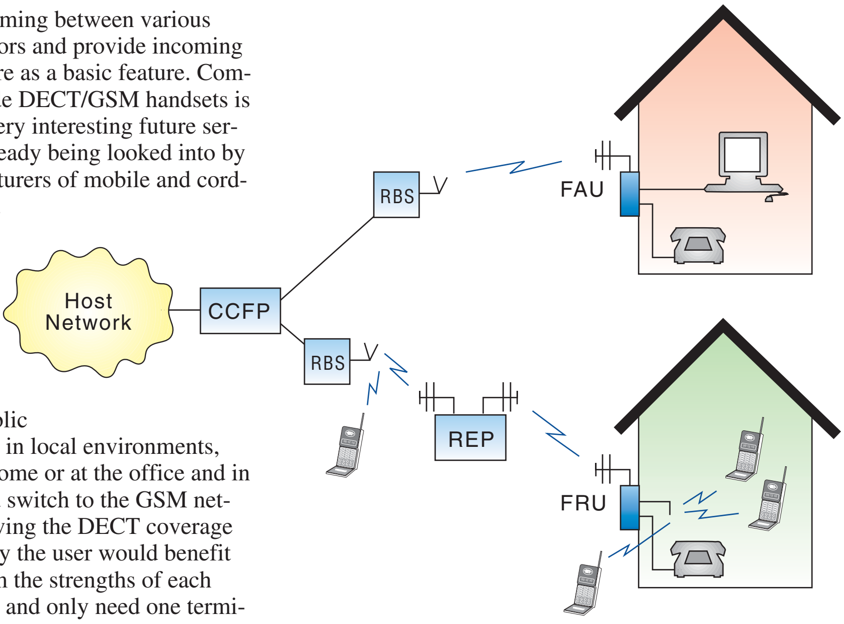


Figure 5 DECT wireless local loop

4 References

1 ETSI. *Radio Equipment and Systems (RES).* Digital European Cordless Telecommunications (DECT) reference document. (ETR 015.)

2 ETSI. Radio Equipment and Systems (RES). *Digital European Cordless Telecommunications (DECT) : a guide to DECT features that influence the traffic capacity and the maintenance of high radio link transmission quality, including the results of simulations.* (ETR 42.)

3 ETSI. Radio Equipment and Systems (RES). *Digital European Cordless Telecommunications (DECT) common interface : services and facilities requirements specification.* (ETR 043.)

4 ETSI. Radio Equipment and Systems (RES). *Digital European Cordless Telecommunications (DECT) system description document.*

5 Dijkstra, S, Owen, F. The case for DECT. *Mobile communications international issue,* 15, 1993.

6 Davies, W. *Pan European mobile communications : from DECT to HIPERLAN : The cordless revolution in office networking.*

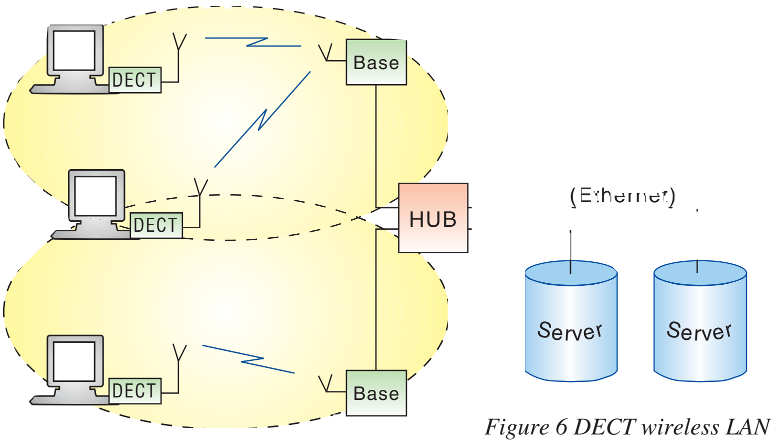


Figure 6 DECT wireless LAN

Satellite based mobile communication – today and in the future

BY ARVID BERTHEAU JOHANNESSEN AND SVEIN ROGER SKALAND

1 Introduction

Public mobile satellite communications was pioneered by Inmarsat, and had its roots in the need for reliable communications for deep sea vessels. Since the introduction of maritime satellite communications, aeronautical and land mobile communication systems have been introduced by Inmarsat as well as other organisations. With the introduction of land mobile satellite communications, the need for lighter portable terminals became apparent. This has resulted in the development of transportable terminals and more recently, briefcase terminals.

Due to the need for global coverage and the constraints of the available satellite technology, the size of the mobile antenna has remained relatively large, and to a certain extent restricted development of smaller mobile units. Mobile satellite terminals have therefore up to this point in time served a niche market, and have complemented rather than competed with terrestrial mobile communication systems.

However, with the introduction of more powerful satellites, handheld satellite terminals become quite feasible, and recently several systems offering handheld communications have been proposed. Is it likely that this evolutional trend will change the complementary relation between satellite and terrestrial mobile systems?

In this article, future mobile satellite systems and its likely development is discussed in relation to current and future terrestrial mobile systems. Emphasis is given on network aspects as this is regarded as an important area for the evolution of mobile satellite systems towards a higher degree of compatibility with the terrestrial mobile systems.

2 Never beyond reach

The history of mobile satellite communication coincides in broad terms with the history of the International Maritime Satellite Organization, Inmarsat, recently changed to the International Mobile Satellite Organization. The need for reliable communications on sea resulted in the founding of Inmarsat in 1979, and the introduction of the Inmarsat-A system in 1982. Inmarsat is an international intergovernmental partnership, where each member country is represented by one signatory. The number of signatories is per today 72 and still increasing. USA, represented by COMSAT, is the largest shareholder with a share of approximately 25 %. Norway, represented by Telenor, has a share of approximately 8 % and is the third largest signatory. Inmarsat offers global mobile satellite communications through four ocean regions served by satellites in the geostationary orbit and some 30 Land Earth Station (LES) sites geographically distributed all over the globe. Several of these LES sites serve more than one ocean region and more than one Inmarsat system. The LESs are owned and operated by signatories, and the Norwegian LES at Eik in the south-west of Norway, is currently the largest LES in terms of paid minutes. Among the manufacturers of LESs and terminals, known as Mobile Earth Stations (MESs) are the Norwegian Nera A/S and the Danish Thrane & Thrane.

The Inmarsat service portfolio includes through five different systems. Inmarsat-A, an analogue system for telephony, telex and data services, is still the main work-horse. Today, there are more than 25,000 terminals out there, mainly installed on-board ships. As Inmarsat-A is one of the very few systems offering global communications, including areas where no other communication is available, the portable terminals are used also for special purposes such as news reporting and disaster relief communication.

The other Inmarsat systems are; Inmarsat-B, the successor of -A, which has many similarities to -A but uses digital modulation. Inmarsat-M is a system for telephony and data using smaller and cheaper terminals than the Inmarsat-B system. Inmarsat-C is a low-speed data message system based on store-and-forward techniques and furthermore Inmarsat-Aero is a system specially designed for aeronautical communications. With the introduction of -M, -C and -Aero, Inmarsat has taken its first steps towards the offering of communications not only at sea, but also on land and in the air.

Table 1 shows an overview of the different Inmarsat systems and their features.

In recent years, several regional mobile satellite systems have been planned and/or introduced providing for more competition and variation of terminal types. Among such systems are AMSC (USA), Optus (Australia), Geostar (USA), Locstar (Europe), Prodat (Europe), Omnitracs (USA) and Eureltracs (Europe). These systems offer a mix of voice and data services.

Some of these systems utilise Ku-band frequencies (12 – 14 GHz) rather than the L-Band frequencies (1.5 – 1.6 GHz) used by most mobile satellite systems. The mentioned systems are aimed at serving regions, such as North America and Europe, and are designed for the special service requirements of those regions. Therefore, some of these systems are providing closed user group services rather than public switched services.

An evolution of the Inmarsat-M and aeronautical systems is currently planned, aiming at further reduction of terminal size and cost. A mobility management scheme is currently under development, and will also provide personal mobility by the use of SIM cards. The systems will be enhanced for later introduction of a selection of ISDN supplementary services.

Table 1 The Inmarsat systems

	-A	-B	-M	-C	-Aero
Services	· telephony · telex · data · fax	· telephony · telex · data · fax	· telephony · data · fax	· telex · data	· telephony · data · fax
Terminal type	· maritime · transportable	· maritime · transportable	· maritime · land mobile · briefcase	· maritime · land mobile · briefcase	· airborne
Year of introduction	1982	1993	1993	1990	1990

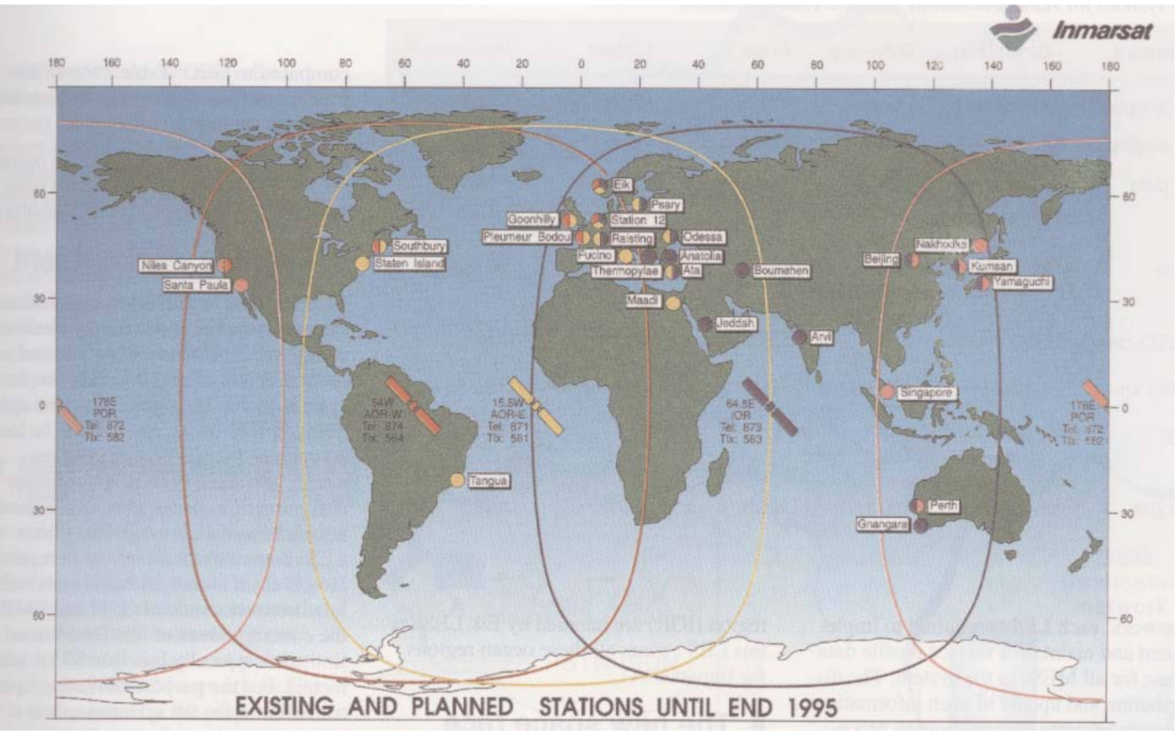


Figure 1 The Inmarsat-A ocean regions and Land Earth Stations

3 The traditional mobile satellite system

Most satellite systems for public or civil applications have so far used Geostationary Earth Orbit (GEO) satellites. Other satellite constellations such as Low Earth Orbit (LEO) and Intermediate Circular Orbit (ICO) have been used in military, navigation and distress and safety systems.

3.1 A typical GEO system configuration

The typical GEO system consists of four main parts:

- The space segment, consisting of the satellites in the geostationary orbit (35786 km height) and their Telemetry and Tracking Control Stations, providing communication between mobile user terminals and the terrestrial gateway
- The user terminal, referred to as the Mobile Earth Station (MES), operating at L-band (1.6/1.5 GHz) for both transmit and receive channels
- The terrestrial gateway, referred to as the Land Earth Station (LES), operating at C-band (6/4 GHz) or Ku-band (12 – 18 GHz) for both transmit and receive channels
- The Network Coordination Station (NCS), providing network control and channel assignment for mobile and fixed originated calls.

Some systems, such as the Inmarsat systems provide global service, while other systems provide national or regional service. Although the number of LESs and MESs vary between these systems, they are all utilising a similar system configuration.

The services provided are mobile extensions of PSTN services, making use of Single Channel Per Carrier (SCPC) mode for the communication links via transparent satellite transponders with global and/or a few spot beams covering large areas on the surface of the earth.

Figure 1 shows the coverage of the Inmarsat-A system with its ocean regions and LESs.

3.2 Access Control and Signalling Equipment

The Access Control and Signalling Equipment (ACSE) is the sub-system of the LES interfacing the Radio Frequency (RF) equipment with the terrestrial networks. The ACSE was typically developed to interwork with terrestrial protocols such as CCITT Signalling Systems No. 5 and R2, and did therefore mainly provide basic telephony services. Some regional systems have been updated to accommodate some ISDN services and associated signalling. The Inmarsat system, being a world-wide system, is somewhat limited in its support of ISDN capabilities due to the lack of global standards being implemented.

The ACSE system is mainly related to call set-up, call maintenance and call clearing functionalities. Roaming in regional systems are simplified due to the centralised control, while roaming in the global Inmarsat system is based on a manual system, requiring the fixed user to have some knowledge of the position of the mobile user. However, systems introduced recently have implemented registration capabilities which could be used for automatic call rerouting purposes. It still remains to implement global rerouting capabilities in the terrestrial network before the full use of such registration databases can be made.

Due to the lack of interrogation capabilities between LESs in the global Inmarsat

Table 2 Proposed systems for handheld mobile satellite communication

	Iridium	Globalstar	Odyssey	Aries	Ellipso	Inmarsat-P
Services	• telephony • paging • data	• telephony • GPS • paging • data	• telephony • GPS • data	• telephony • data • paging	• telephony • data • paging	• telephony • data • paging
No. of satellites	66	48	12	48	24	10
Orbit type	LEO circular	LEO circular	ICO	LEO circular	elliptical	ICO
Orbit height	765 km	1389 km	10,355 km	1020 km	429–2903 km	10,355 km
Orbit inclination	87 °	52 °	55 °	90 °	64 °	45 °
Planned operational	1998	1999	1997	1998	1997	1999

network, each LES is required to implement and maintain a service profile database for all MESs in the system. The distribution and update of such information require separate commissioning procedures to be implemented.

3.3 Service Provision

The provision of the service to the mobile user is probably where the current satellite systems differ the most. In the case of most regional systems, the system owners both operate the system and provide the service to the end users. The mobile user do in most cases pay a subscription fee.

Inmarsat provides the service to the end user through the operators of the LESs. The service is available to the end user through several operators in the same ocean region, competing with each other. For mobile-to-fixed calls, the mobile user is free to choose the service provider with the best deal for him amongst all the providers in the particular region. There is no subscription fee for the mobile user, and when entered into the system as a user, he can freely choose his service provider on a per call basis. The call charges will be billed to the customer by one LES operator, depending of the MES nationality.

Fixed-to-mobile calls are routed via a specific LES in the ocean region chosen based on routing agreements of the network operator in the country where the call is originated. This means that all calls initiated in Norway to Inmarsat-A MESs operating in either Atlantic Ocean Region East (AOR-E), Atlantic Ocean Region West (AOR-W) or Indian Ocean region (IOR) are handled by Eik LES (as this LES covers all these ocean regions for Inmarsat-A).

4 The new space race

The evolution seen in satellite communications recently is based on the wish to combine the main advantages of satellite systems and terrestrial based mobile cellular systems, namely global coverage and small terminal size, into one ultimate satellite communication system for voice and low rate data communication. A large number of companies and organisations have declared their plans to realise such a system. Some of these systems are described as “cellular extensions”, providing communication where cellular coverage is not available or profitable, but most of the systems are announced to have global coverage. The operation via “lower-than-geostationary” orbits is also a common denominator of these systems. The use of LEO or ICO orbits significantly reduces the transmission delay and signal loss as compared to GEO. Table 2 shows an overview of some proposed systems and their properties [1].

Some of the current GEO system operators have also recently announced plans for handheld services via GEO satellites.

5 Mobility management

The traditional satellite communications system based on geostationary satellites using global beams is, from a ground segment point of view, a fairly simple system. Actually, a satellite system operating in four ocean regions may be looked upon as four communication networks, and there is no need for further definition of location areas within these networks as the coverage area of one LES is equal to the entire ocean region. However, in terrestrial based mobile cellular networks such as GSM and NMT, the coverage areas of one Base Station is limited to typically less than 50 km (diameter). For the purpose of routing fixed-to-mobile calls, the system needs to know at any time where to find the mobile terminal and how to route calls to this location. Such functionality is referred to as mobility management.

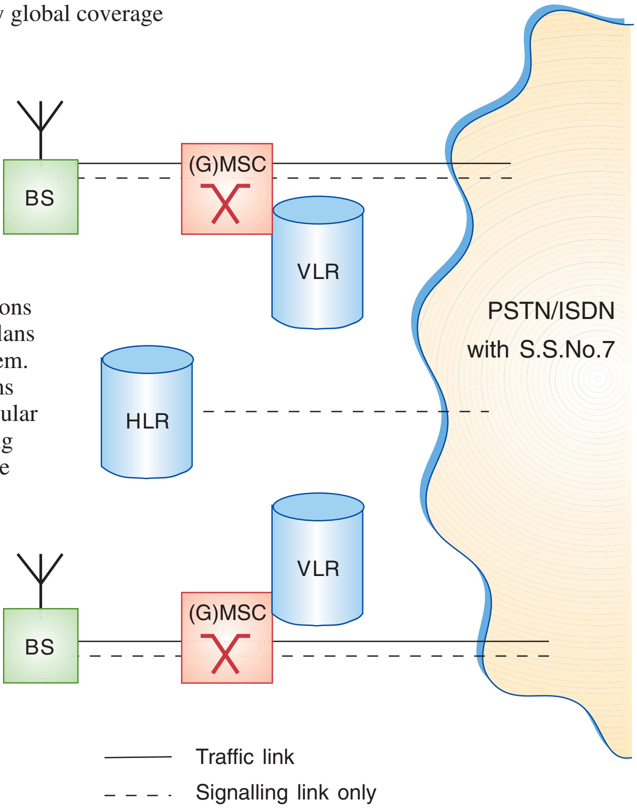


Figure 2 GSM network elements and interconnections

As satellite communications evolve, mobility management will become a necessary feature also for these systems. The next generation Inmarsat geostationary satellites, Inmarsat-3, feature spot-beam operation. Existing services use global beam signalling, but future services may use spot beam signalling and this will decrease the size of the location areas. Also, the use of lower orbital configurations results in reduction of the coverage areas and the need for mobility management functionality similar to that of terrestrial mobile cellular systems will emerge.

5.1 Mobility management in GSM

One of the best defined technologies for mobility management is the one used in the European cellular system GSM. The GSM network includes the main switching functions as well as the data bases needed for subscriber data and mobility management. The GSM network architecture is used as basis for the proposed satellite system network architecture. Figure 2 shows the relevant GSM network elements and their interconnections [2].

BS The Base Station (BS) represents the transition element between the fixed and the mobile environment and basically contains the RF parts of the system.

MSC The Mobile services Switching Centre (MSC) acts mainly as an ISDN exchange and represents the interface between the fixed and the mobile environments. Mobility management functions, such as call routing for both mobile originated and mobile terminated calls and activation of handover procedures, are mandatory for an MSC.

HLR A Home Location Register (HLR) is essentially a database containing location and subscriber data for the Mobile Stations (MSs) associated with a specific network (home network). The HLR is typically a stand-alone element without switching capabilities.

VLR A Visitor Location Register (VLR) is related to one or more location area(s) and contains information about the MSs currently located in this area. The VLR is always integrated with each MSC.

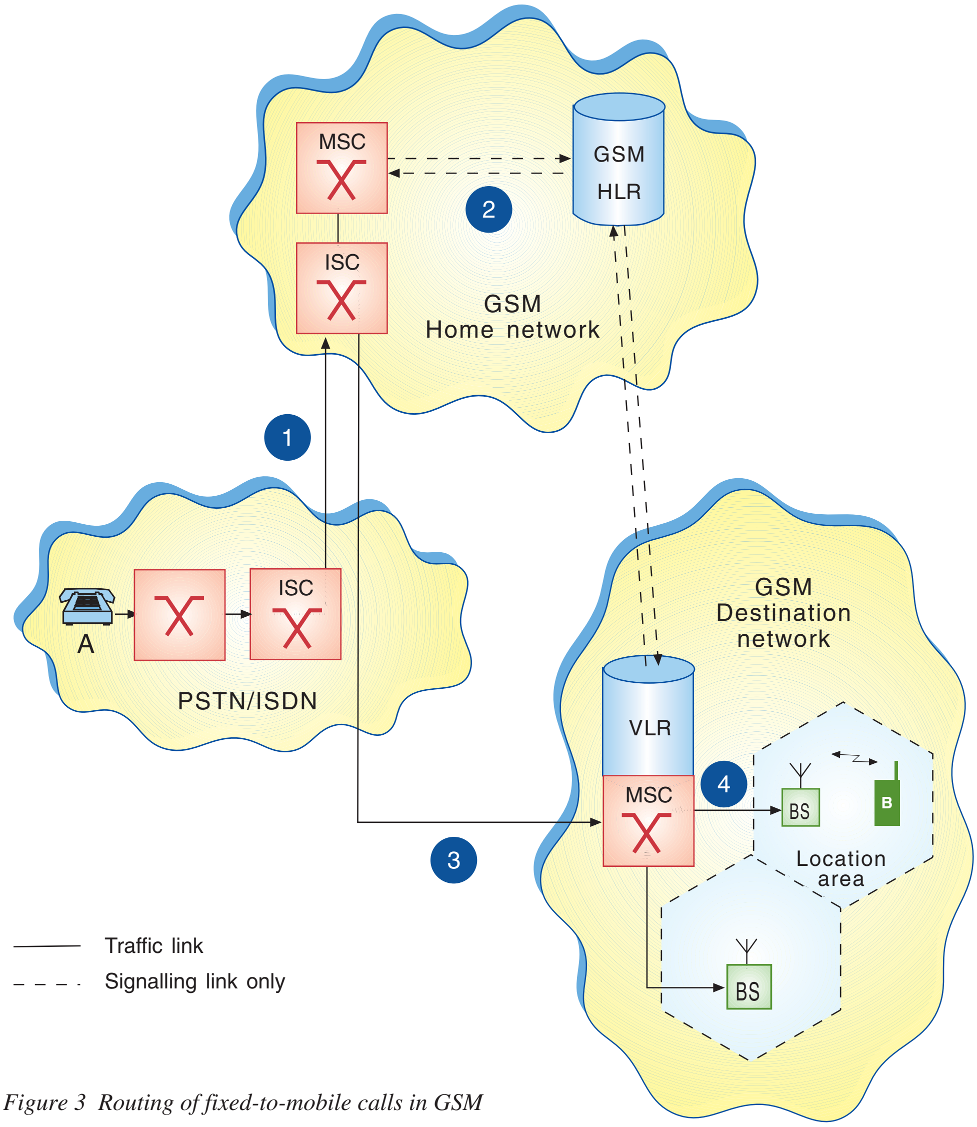


Figure 3 Routing of fixed-to-mobile calls in GSM

GMSC In order to set up a call towards a GSM user, the call is first routed to a gateway switch, referred to as a Gateway Mobile services Switching Centre (GMSC). The GMSC is in charge of obtaining the necessary location information and of routing the call to the appropriate destination MSC. The GMSC is in principle an independent element or a function integrated in a digital telephony switch, but currently the GMSC is implemented as a function in the MSC.

The directory number of a GSM Mobile Station (MS), the MS ISDN number, is used by a calling subscriber for identification of the called MS. The structure of the MS ISDN number is according to CCITT Rec. E.164. E.164 is the CCITT numbering plan for the ISDN era.

The E.164 structure is

MS ISDN = CC + NDC + SN

where

CC = Country Code
NDC = National Destination Code
SN = Subscriber Number

The MS ISDN number looks like a standard PSTN/ ISDN number, but the knowledge of the NDC identifies an operator within a country and not an area code.

Routing of a call from a fixed subscriber A to a mobile subscriber B, which has roamed outside his home GSM network, is described below and illustrated in Figure 3. Essential for this routing principle is the assignment of a roaming number. This roaming number is part of the national numbering plan of the destina-

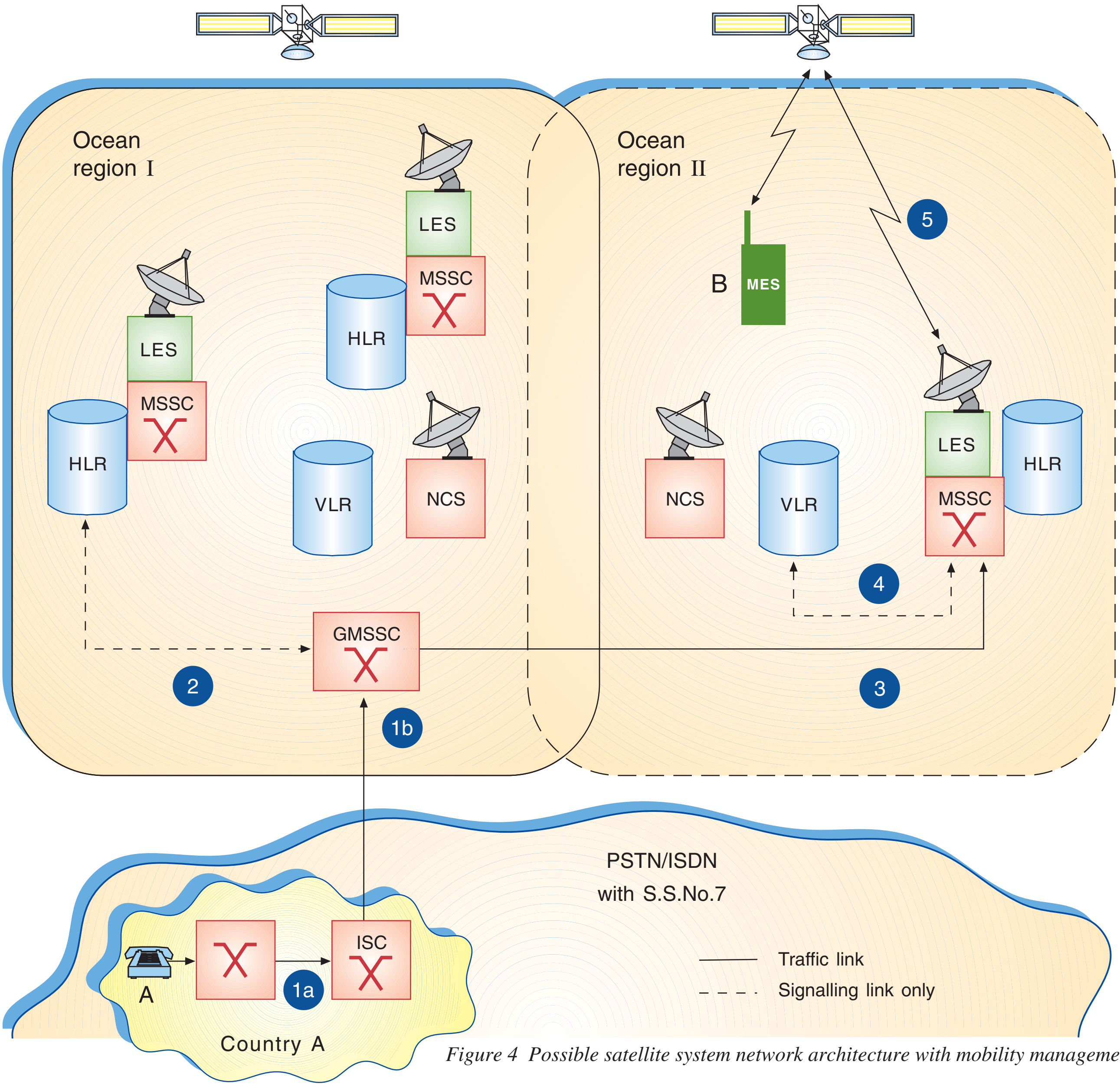


Figure 4 Possible satellite system network architecture with mobility management

tion country and will be temporarily assigned to the mobile subscriber B. The numbers 1–4 correspond with the numbering in the figure.

1 The call is routed to an MSC in the home GSM network of B based on the CC + NDC of the MS ISDN number.

2 The MSC interrogates the appropriate GSM HLR, based on the first digits following the NDC, via an SS No. 7 link to obtain an E.164 roaming number. This roaming number is either:

- kept in the GSM HLR, or
- downloaded from the destination GSM VLR on a per call basis

3 The call is routed to the destination MSC in the destination GSM network using the roaming number.

4 The destination MSC will page for B in the destination location area using location information kept in the GSM VLR and the call will be established from A to B.

5.2 Mobility management in a satellite communication system

For the routing of fixed-to-mobile calls in a satellite system, similar network elements and procedures as used in GSM may be used. Figure 4 shows a possible architecture for a satellite system with mobility management [3], [4]. In this example, the ocean regions serve as location areas. For simplicity, only two ocean regions are shown.

The overall structure of the *Inmarsat mobile international number* is:

CCC T X_1 ... X_k

where

CCC	is a three digit country code allocated to Inmarsat
T	is the T digit identifying the different systems (e.g. -M, -B etc.)

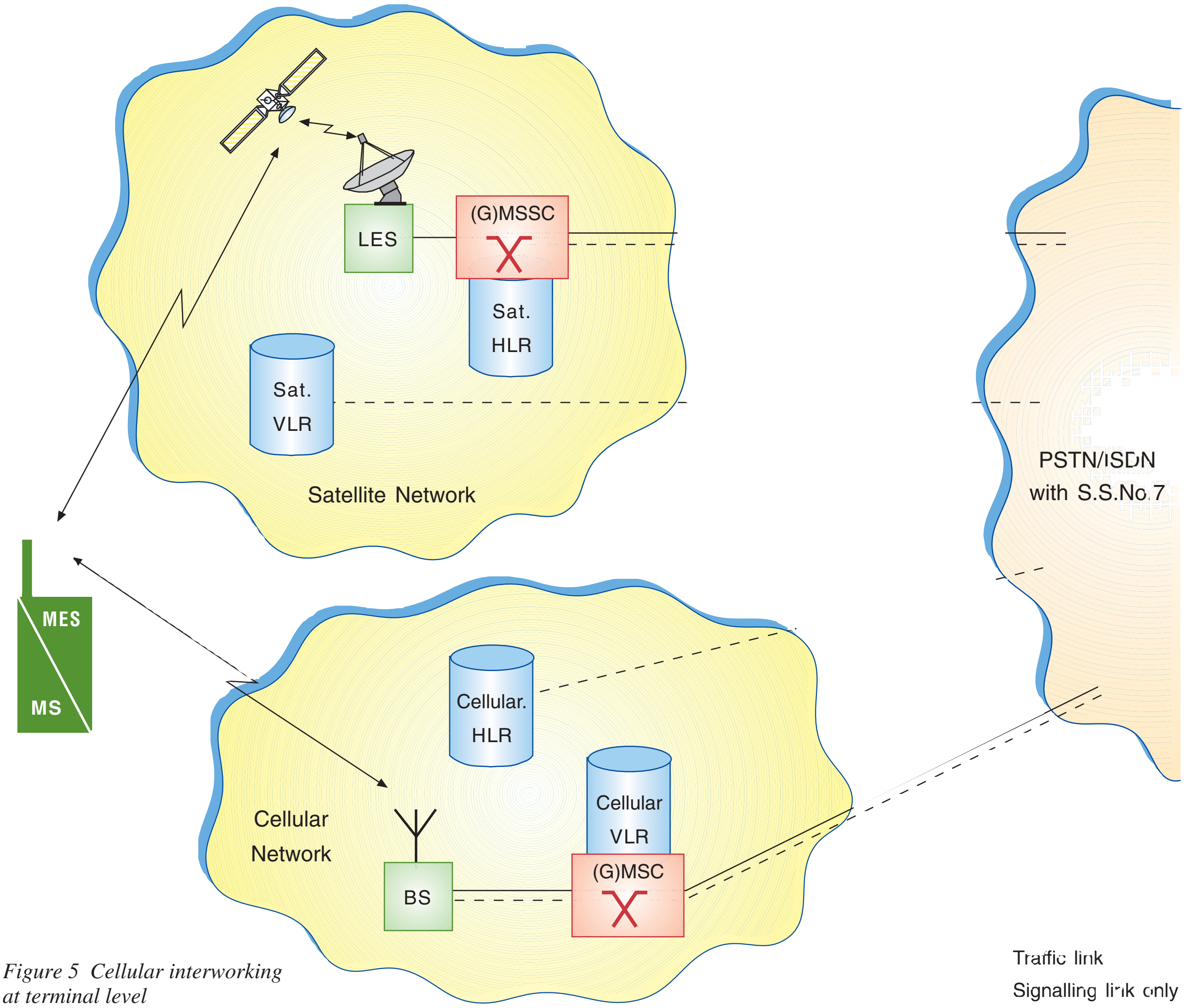


Figure 5 Cellular interworking at terminal level

X_1 ... X_k is the MES identifier, including info. about signatory subscribed to.

For Inmarsat, it should be noted that the country code is not related to a country, but identifies an ocean region within the Inmarsat system. This omits the need for routing calls to a roaming user via the user's home country, known as "tromboning" in GSM.

Routing of a call from a fixed subscriber A to a mobile subscriber B, could be done in the following way (see Figure 4):

1a The call is routed to the originating country's International Switching Centre (ISC) based on the international prefix (e.g. 095 in Norway, 00 in the UK)

1b The call is further routed to a Gateway Mobile Satellite Switching Centre (GMSSC) based on the CCC and T digit

2 The GMSSC interrogates the appropriate HLR, via an SS No. 7 link. The appropriate HLR is chosen based on the subscription information in the MES identifier. The HLR supplies information about the VLR where the MES is currently logged on, and also an E.164 roaming number is provided to the interrogation node (GMSSC). This roaming number is either:
· kept in the HLR, or

· downloaded from the destination VLR on a per call basis

3 The call is routed to the destination MSSC/LES using the roaming number.

4 The MSSC interrogates the VLR for more detailed location information (e.g. spot beam) and subscriber data

5 The call is established from A to B.

Unlike the GSM system where roaming involves a limited number of network operators, global satellite systems, like the Inmarsat systems, will require global roaming and routing capabilities in the fixed network. This will require standardised mobility management specifications at an international level.

6 The first steps towards cellular interworking

The introduction of mobility management in satellite systems not only automates the routing of fixed-to-mobile calls, it is also a first step towards a higher degree of compatibility between satellite systems and other networks like PSTN/ISDN and terrestrial cellular systems. The possible use of standardised network elements and interfaces enables

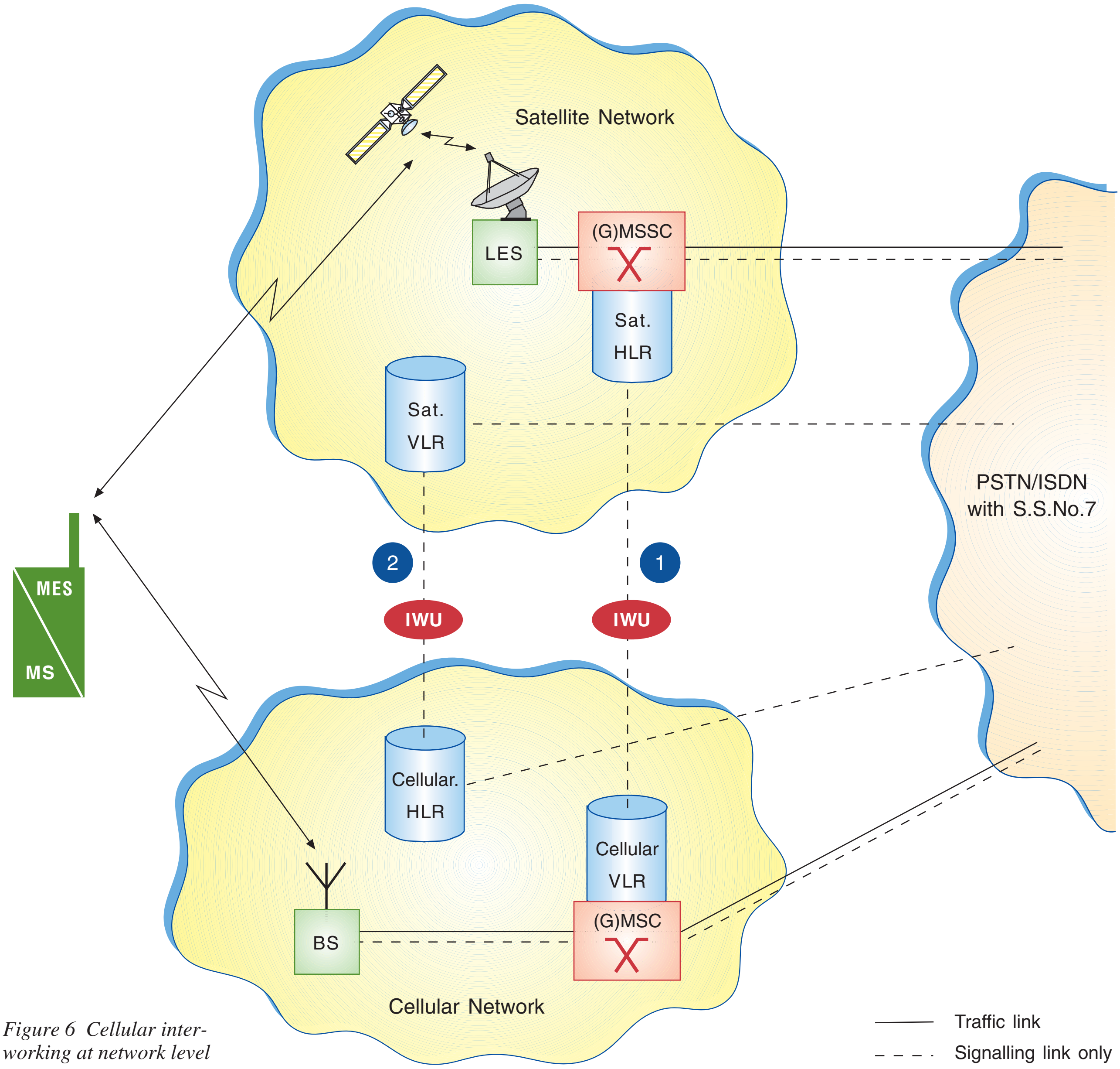


Figure 6 Cellular interworking at network level

for the possibilities of interconnecting satellite systems like Inmarsat and cellular systems like GSM. Using what has been named a dual-mode terminal, the user should be able to roam not only within a system, but also across different systems. Cellular interworking could be implemented at different levels as discussed in this section [5].

6.1 Cellular interworking at terminal level

Figure 5 shows how a satellite system could be integrated with a cellular system at a terminal level. This interworking scenario does not include any integration between the satellite system and the cellular system at a network level except possibly a call forwarding capability for the transfer of calls between the two networks. The dual-mode terminal will include all functionality to function as a cellular terminal of a given standard and all functionality to function as a satellite terminal. In addition, the terminal must have functionality to switch between satellite mode and cellular mode either manually or automatically. All calls to the dual-mode terminal in satellite mode will be routed via the satellite network and all calls to the terminal in cellular mode will be routed via the cellular network. Location updating in the satellite network will be independent of location updating in the cellular network.

6.2 Cellular interworking at a network level

Figure 6 shows how a satellite system could be integrated with a cellular system at a network level. Signalling links between the location registers of the satellite network and a cellular network allow for location updating and roaming across the systems. Three different scenarios for interworking at a network level are identified. These three scenarios are all illustrated in Figure 6.

Scenario 1:
Satellite system with cellular extension
This could be an attractive service for users who are primarily satellite communication users, wanting cellular coverage when visiting areas where this is available and preferable. The satellite network HLR interworks with the VLRs in the cellular network, as illustrated by signalling link 1 in Figure 6. The Interworking Unit (IWU) is required in order to provide for signalling compatibility between the systems. It is likely to assume that the satellite system HLR should look like a cellular HLR seen from the cellular network.

Scenario 2:
Satellite system as an extension to a cellular network
This could be an attractive service for users who are primarily cellular users, wanting communication when visiting areas outside cellular coverage. The satellite network VLRs interwork with the HLR in the cellular network, as illustrated by signalling link 2 in Figure 6. As in scenario 1, an IWU is required in order to provide for signalling compatibility between the systems. It is likely to assume that the satellite system VLR should look like a cellular VLR seen from the cellular network.

Scenario 3:
Combined configuration
The two scenarios above could very well co-exist to serve all categories of users. Both signalling links 1 and 2 in Figure 6 are then required.

It should be noted that the implementation of cellular interworking at a network level requires extensive investigations on operational and regulatory aspects, and will require interworking between several cellular operators and the satellite system operator. This also includes commercial arrangements with respect to subscriptions including accounting and billing. For a global satellite system, this is further complicated as interworking may be required with several regional cellular standards.

7 Two worlds meet

Until now, terrestrial based mobile communications and satellite communications have evolved more or less on individual basis. The first generation terrestrial based systems could be illustrated by NMT, starting out as a Nordic system but soon adapted by more nations such as the Netherlands and Switzerland. The need for a pan-European system emerged into the GSM system specified through European standardisation organisations like ETSI and recognised as a second generation cellular system. Other second generation systems with many similarities to GSM are the Japanese PDC system and the North American DAMPS system.

An evolutional trend seen among cellular systems is the utilisation of higher frequency bands, and smaller cell size for increased frequency reuse. These systems, usually referred to as PCS or PCN systems, are characterised by the use of second generation technology or a subset of these standards.

Within satellite communications, we have seen the evolution towards smaller terminals, operation both on sea, land and in the air, and smaller cell areas as introduced by spot beam operation and/or lower orbital configurations. As the traffic capacity of the satellites increases, the global coverage remains. The ground segment network capabilities evolve towards a higher degree of compatibility with terrestrial mobile systems.

The next step within the evolution of mobile communications is the specification of the third generation system FPLMTS (Future Public Land Mobile Telecommunication Systems) initiated by the international standardisation organisation ITU. The definition of FPLMTS represents an important step in the direction of merging the two “worlds”, as it is decided that FPLMTS should contain satellite components. This means that both satellite and terrestrial links should be used to provide for a wide range of services within one global mobile communications system. However, the recent deregulation of mobile communication and introduction of competition among operators and service providers, have radically changed the standardisation process. The acceptable implementation time has been reduced and the probability to join forces in a long term development of a new system seems less likely. This may lead to a situation where cellular operators prefer to further develop their existing systems, rather than embarking on the development of a completely new system.

At the same time, although in a smaller scale, the mobile satellite operators have experienced a similar development. Introduction of more competition has reduced the likelihood of one global standard. This picture, together with the increasing number of cellular systems to interwork with, further complicates the integration process. In such environments, mobile satellite communications and terrestrial mobile communications could still share the rapidly growing market of mobile communications as complementary rather than competing systems, although not as an integrated system on a world-wide scale.

References

1 Crosbie, D M. The new space race : satellite mobile communications. *IEE review,* May, 111-114, 1993.

2 Mouly, M, Pautet, M-B. *The GSM system for mobile communications.* Published by the authors, Palaiseau, 1992.

3 Sengupta, J R, Skaland, S R. Mobility management in Inmarsat mobile satellite communications systems. In: *2nd international conference on universal personal communications, ICUPC’93,* Ottawa, 1993.

4 Johannessen, A B, Ytterbø, P A. *Inmarsat-P mobility requirements.* Kjeller, Norwegian Telecom Research, 1993. (TF-report R 6/93.)

5 Ytterbø, P A, Johannessen, A B. *Inmarsat-M : cellular interworking.* Kjeller, Norwegian Telecom Research, 1994. (TF-report R 17/94.)

Cellular systems for mobile data

BY STIG KASPERSEN

1 Introduction

This article describes some technical aspects of the data services which are available from Telenor Mobil.

There are two main categories of systems capable of mobile data communication; dedicated data systems and shared systems for voice and data. Amongst the latter, some systems are designed for voice with no specific system facilities for data, as is the case for the analogue Nordic Mobile Telephone System (NMT), while others, for instance GSM, are designed as general purpose systems.

When designing communication networks, voice and data have different and sometimes contradictory requirements (Table 1), so when integrating both in a single network, as done in GSM, some compromises have been made, but not at the cost of voice telephony which is the main service of GSM.

Dedicated data systems do not need to compromise with voice, but even so there is a variety of contradictory requirements for different kinds of data transmission.

Telenor Mobil operates four systems capable of two-way data communication. These systems are

- NMT (two systems, NMT450 and NMT900)
- GSM
- Mobitex®[1]

Each system has different qualities, thus, they attract different customer needs. NMT-450, for instance, has by far the best geographical coverage in rural areas of Norway and in the surrounding oceans. Mobitex only covers the most populated areas, but is well suited for transmission of short data blocks between a fixed host computer and a fleet of mobile terminals. GSM is the superior system for transfer of bulk data, for instance fax or large data files.

2 NMT-450 and NMT-900

In order to use NMT for data communication one has to connect a telephone modem to the NMT telephone. This requires an NMT telephone which is equipped with a line interface.

In principle, NMT gives the same possibility for data communication as does the ordinary telephone networks, the PSTNs. However, NMT suffers from some disadvantages compared to PSTN:

- When a mobile telephone is moving, the radio channel quality heavily varies. As an average the quality is lower than that of a PSTN connection.
- When moving from one radio cell to another there will be a notable break in communication while switching between cells ('handover'). This will cause loss of carrier between the modems. Most of the telephone modems disconnect immediately when loosing the carrier, but some modems have an adjustable timer making it possible to delay disconnection. In order to reduce the handover problem it is possible to disable the handover function for users who subscribe to this.

Table 1 Contradictions between voice and data requirements in a digital communication system

Requirements if optimising for data	Requirements if optimising for voice
Moderate time delays are acceptable. Maximum acceptable delay depends on the user application.	Minimised, constant time delays
Data networks strive for higher bit rate and as the available bit rate increases new applications evolve.	Using the bit rate necessary for acceptable voice quality. As the technology evolves the bit rate required decreases.
Wide range of session lengths and data volumes. Average amount of information involved is much smaller than for voice.	The session length (and amount of information) is relatively uniform (in the order of minutes).
Very low bit error rate required. A retransmission strategy is often used to achieve this.	Relatively high bit error rates are acceptable without quality degradation. Data are only transferred once.
Long set up times are undesirable, especially for small sessions.	Set up times of a few seconds are acceptable for voice telephony. For 'Trunked Radio Systems' (e.g. systems based on the industry standard MPT-1327 or the ETSI TETRA standard), where normal sessions often are in the magnitude of a few seconds, set up times less than a second is required.
Asymmetrical sessions for most applications – data is transmitted in one direction at the time.	Symmetrical sessions – same capacity made available in both directions.
Data networks are implemented with packet switching.	Voice networks are implemented with circuit switching.
Designed for efficient communication between a host computer and group of remote, mobile terminals.	Mobile telephone systems are designed for efficient communication between two parties. 'Trunked Radio Systems' are designed for efficient communication between a base operator position to a fleet of mobile positions.
Efficient use of spectrum and transmission resources when used for data traffic of 'bursty' nature.	Efficient use of spectrum and transmission resources when used for telephony voice.

[1] *Mobitex is a registered trademark from Telia Mobitel.*

To obtain acceptable data communication over NMT one should as a general rule:

- communicate from a spot with strong radio signal from the base station
- avoid data sessions with a moving NMT terminal
- subscribe to the 'handover disabled' function
- utilise error correction protocols if possible.

2.1 Data

Modern telephone modems support several protocols enabling the users to select different speeds, error correction and compression. Since both sides in a session have to support the preferred set-up selection, the lack of functions supported at one end may be a limitation.

The radio channel quality normally sets limits to a data session: With excellent quality and a stationary NMT telephone a data bit rate of 14,400 bit/s with no error correction is possible. Under poor or heavily varying conditions the maximum usable bitrate decreases significantly and error correction protocols will be needed in order to obtain reliable data transfer.

2.2 Fax

Modern fax machines may select speeds between 2400 bit/s and 9600 bit/s. At the beginning of a session the fax machines negotiate about speed and protocols. With good radio channel quality a bit rate of 9600 bit/s is possible. However, if the quality is excellent at the start of a session, thus resulting in a high data bit rate, and then reduced because the NMT moves away from the good radio coverage, the data speed will not be reduced. This will result in an increased bit error rate and reduced transmission quality.

Handover between base stations will cause loss of carrier and can be seen as lost lines in the transmitted documents if no error correction is used. A normal handover does not cause much harm, but a poor handover may seriously damage the transmitted document or may even cause disconnection of the line.

Some fax machines support the ECM error correction protocol. If ECM is used the data to be transmitted is grouped into frames (64 or 256 bits) and blocks (256 frames) together with error detection information. If errors are detected in a block the receiver requests retransmission of the erroneous frames.

3 Data in GSM

While NMT is a system for mobile voice telephony, GSM is a multi service system designed to support communication of various types. From the start the system has been designed to offer many data services. Basically, the services which are provided to users of ISDN (Integrated Services Digital Network) or PSTN (Public Switched Telephone Network) have been included as far as the limitations due to radio transmission allow. In addition, GSM provides an untraditional set of data services not found in the fixed networks; the Short Message Service. This is a packet switched service, using the GSM signalling channels for transmission of limited sized text messages.

3.1 Circuit switched data services

3.1.1 General

The purpose of the GSM data services is to provide communication through the GSM network to some end user equipment commonly used in fixed networks, such as faxes or PCs. As the implementation of new functionality in the network as well as in the mobile stations is an ongoing process, new services and facilities, and enhancement of the existing ones, are continuously made available to the GSM subscribers.

3.1.2 Data services offered

The functional part of a GSM mobile station which performs the adaptation between a specific terminal equipment (e.g. a PC) and the generic radio transmission part, is called TAF (Terminal Adaptation Function). The adaptation function at the other end, between GSM and the external networks, is called the IWF (Network Interworking Function).

TAF and IWF are entry points into GSM. Their functions depend on the type of end-to-end service to be supported. If, for instance, fax is to be used, the TAF and the IWF involved are instructed to perform the required adaptation functions for fax communication during the session. The GSM network between TAF and IWF is not concerned with the end-to-end service, only with the transport of the corresponding data flow through GSM.

GSM offers a set of so-called 'bearer services' to provide data communication between TAF and IWF. The gross bit rate between TAF and IWF is 12,000 bit/s, but different levels of Forward Error Correction are added, thus resulting in available bit rates of 1200 bit/s, 2400 bit/s, 4800 bit/s or 9600 bit/s. The remaining bits are used for error correction, so that the lower the bit rate, the more powerful is the error correction.

Due to the varying noise conditions on the radio path and the wish of the GSM designers to offer a pre-defined quality of

Some milestones in the history of mobile data services from Telenor

1981	The automatic cellular system NMT-450 was introduced. The two NMT systems in operation today do not offer data services, but data communication is possible by use of analogue telephone modems. Along with the NMT growth better user equipment for data has been made available to the market and a few users with special needs are using NMT for data.
1985	Trials started with a proprietary mobile data system in Oslo. Dual mode mobile stations for NMT-450 and mobile data were used.
1989	A mobile data service based on the Mobitex® technology from Ericsson was introduced, utilising a gross data rate of 1200 bit/s on 2 x 25 kHz radio channels.
1993	Mobitex® was upgraded with second generation technology, increasing the gross data rate to 8000 bit/s on 2 x 12.5 kHz radio channels.
1993	Trials with GSM data started.
1994	The Short Message Service was introduced as an ordinary GSM service.
1995	Data and fax was introduced as ordinary GSM services.

Table 2 Some characteristics of transparent and non transparent data services in GSM

Qualities	Transparent	Non Transparent
Throughput	Constant	Variable
Delay	Constant	Variable
Bit error rate	Variable dependent of noise conditions on the radio channel and the selected data bit rate	Almost constant. Virtually error free.
Error correction	Forward Error Correction protocol	ARQ (Automatic Repeat Request protocol) together with Forward Error Correction

the data services, two principle types of data transmission are introduced, transparent and non-transparent. A transparent mode connection provides a fixed delay, and error correction through the Forward Error Correction scheme, which means that a constant amount of the bit flow is used for error correction. In this mode there remains a small residual error rate depending on the data bit rate used and the noise conditions on the radio channel. In non-transparent mode a virtually error free data channel is provided by the Radio Link Protocol (RLP). The RLP packetises the data and applies a CRC check (CRC = Cyclic Redundancy Control) after application of the Forward Error Correction scheme. The CRC check indicates to the receiver if errors have occurred, but it does not correct them. If a packet is correctly received it is acknowledged by the receiver and if not, the receiver requests retransmission of the data.

3.1.3 Connections to other networks

3.1.3.1 Connections with users of PSTN (Public Switched Telephone Network)

Ordinary telephone networks, or PSTNs, are widely used for data transmission with analogue audio modems. Since voice transmission over the radio path in GSM is based on coding algorithms which are optimised for speech, standard analogue modem signals are unsuited for coding by the same algorithms.

In GSM the mobile user does not provide a modem for data communication since the transmission through GSM is digital. At the point of interworking between GSM and PSTN the data are transformed between the digital GSM standard and the corresponding analogue modem standard. GSM copes with the most widely used modem standards including the standards specific to Group 3 Fax.

3.1.3.2 Connections with users of ISDN (Integrated Services Digital Network)

The maximum bit rate used for GSM data is 9600 bit/s while the basic bit rate in ISDN is 64,000 bit/s. In order to cope with the provision of ISDN data services to PSTN users, ISDN has standardised digital formats for lower bit rates. This makes possible a way of interworking where GSM and ISDN are considering each other the same way as they do with PSTN. Interworking can then be done using the standardised ISDN digital formats developed for PSTN adaptation.

Since no modems are involved, one of the most notable benefits to the users, when interworking with ISDN instead of PSTN, will be a significant reduction in the session set-up time.

3.1.3.3 Connection with users of PSPDN (Public Switched Packet Data Networks)

Packet switched networks, PSPDNs, such as Telenor's Datapak may be accessed through PSTN via a PAD (Packet Assembler Dissembler for access via asynchronous modems). To utilise this the GSM user needs a subscription to PAD access in a PSPDN network. The connection has to be set up in two steps, first by establishing the connection to the PAD and then addressing the PSPDN subscriber. It is only possible with calls set up by the GSM mobile station.

3.1.4 Requirements to users of GSM data services

GSM fax and GSM data have to be subscribed to separately. A data subscriber is assigned to a telephone number dedicated to data as is a fax subscriber assigned to a telephone number dedicated to fax. This means that a GSM subscriber with a data subscription and a fax subscription, in addition to the standard speech subscription, has three telephone numbers assigned to his SIM-card. This is done in order for the mobile station to separate between incoming speech calls, data calls or fax calls.

At the mobile side the GSM data or GSM fax users need the functional parts as follows:

1 specific external terminal equipment (e.g. portable PC or fax)

2 generic radio transmission equipment (i.e. GSM telephone)

3 terminal adaptation equipment between the terminal and radio parts (e.g. GSM PCMCIA card)

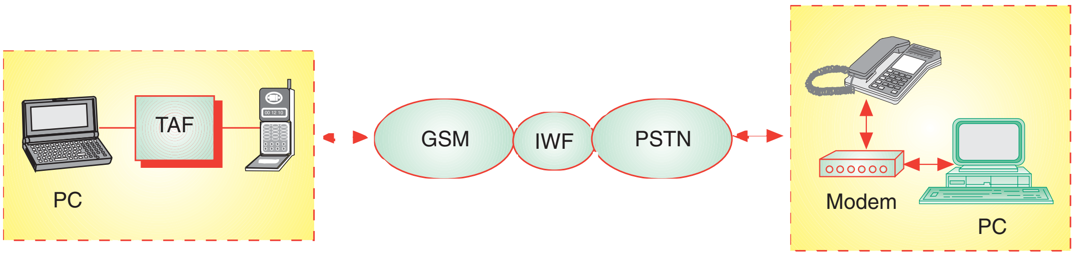


Figure 1 Data communication via GSM

Table 3 Elements in Short Messages particular to submission and reception

Element	Function
Validity Period	Indication of the time period for which the Service Centre shall store an undelivered message.
Service Centre Time Stamp	Information of the time of arrival of a short message to the Service Centre.
Protocol Identifier	Indication of higher level protocols being used or indication of interworking with a certain type of telematic device.
More Messages To Send	Information from the Service Centre to a mobile station whether more messages are waiting to be delivered.
Priority	Indication to the GSM network if the message is a priority message, thus activating specific priority delivery procedures
Messages Waiting	Indication to the GSM network, in case of an unsuccessful delivery, that a message for the recipient mobile station is waiting in the originating Service Centre.
Alert Service Centre	Information from the GSM network to the Service Centre that a mobile station, to which a delivery attempt has failed due to exceeded memory capacity or that it is unreachable, is again ready to receive short messages.

The terminal adaptation equipment available is specific to the GSM telephone, so that a GSM telephone with associated terminal adaptation equipment is required. At the terminal end a standard interface is used. This may be an analogue Group 3 Fax interface – or more common – a PCMCIA interface (PCMCIA = Personal Computer Memory Card International Association). The terminal equipment may then be a PC with a PCMCIA slot and standard programs for fax or data communication – or a digital fax with PCMCIA slot.

For incoming faxes it may be convenient to use the ‘Mobilfax’ store and forward service, which is offered by Telenor Mobil, in combination with GSM. It is then possible to have the faxes directed to ‘Mobilfax’. The subscriber is notified about incoming faxes via a GSM Short Message or by other means, offered by the ‘Mobilfax’ service. The received fax may then be forwarded to the GSM fax or to another fax address at the choice of the user.

3.2 Packet switched data services in GSM – the Short Message Service (SMS)

3.2.1 General

The Point-to-Point Short Message Service enables GSM users to send or receive short data messages. Another Short Message is the Cell Broadcast message, enabling short data messages of general nature to be broadcast at regular intervals to all subscribers in a given geographical area. At present, Cell Broadcast is not used by Telenor Mobil.

A Short Message contains up to 140 octets of user data, normally represented by 7 bit ASCII text characters, corresponding to 160 characters. A received Short Message may be stored in the telephone memory or in the SIM-card.

Point-to-Point Short Messages may be delivered to GSM telephones during an ongoing call as well as in idle state while Cell Broadcast messages may only be received in idle state.

3.2.2 Services available

The Point-to-Point Short Message service distinguishes between the two basic services:

- SM MT (Short Message Mobile Terminated), which denotes the capability to transfer a Short Message from a Short Message Service Centre to one GSM mobile station, and to provide information about the delivery of the Short Message.
- SM MO (Short Message Mobile Originated), which denotes the capability to transfer a short message from a GSM mobile station to a Short Message Service Centre, and to provide information about the delivery of the Short Message.

Short Messages may be used together with GSM telephones with external terminal equipment, but this is not the common use today, since the Short Message functions are well fitted for integration into the GSM telephone stations. Most of the GSM telephones on the market today support receiving of short messages (SM MT) and some GSM mobile stations are also able to send short messages (SM MO). At present, Mobile Originated Short Messages may only be conveyed to other GSM subscribers, but the possibilities may well be extended in the near future.

The originator of a short message may be notified about the status of the short message being sent, for instance[2]:

- when a message is successfully delivered to the addressee
- when a message is stored because the addressee is temporarily unreachable
- when the validity period of an undelivered Short Message has expired
- when a message is rejected by the network for some reason.

3.2.3 Transfer of Point-to-Point Short Messages

The process of delivering a Short Message is a two-stage process; the message is first sent from the message originator to the Short Message Service Centre, which then sends it to the recipient. The recipient of a short message is identified by their telephone number.

The transportation of Short Messages between GSM mobile stations and the Service Centre is included in the GSM specifications, but the Service Centre functionality is not standardised in the GSM specifications, except for the functionality related to the Short Message Service between the Service Centre and GSM mobile stations. For this reason and since no similar services are defined in fixed networks, parts of the service, as seen by the user, are not standardised, but

[2] *The set of notifications offered as default by Telenor Mobil may differ from what is offered by other network operators.*

are specific to the Service Centre operators and manufacturers[3].

Transportation of Short Messages through GSM is carried out on the signalling channels. The signalling system in GSM may be considered as an overlaying packet switched network, adding to the Short Message Service some of the qualities of packet switching.

On the top of the protocol hierarchy used for Short Message transmission there is a transport protocol for short messages, the SM-TP. This is offering an end-to-end service between the GSM telephone and the Service Centre, relying on several underlying protocols within the GSM. There are defined a couple of service elements, for information between entities involved in submission and reception of Short Messages. These elements are listed and the functions explained in Table 3.

3.2.3.1 Mobile Originated Short Messages

In addition to its contents a Short Message must contain the identification address of the final destination and the Service Centre address. When transmission is requested there will be created a single transport layer message, the 'SM-TP SMS-SUBMIT'. This message will then be transferred on various interfaces, using the capabilities of the lower layer protocols. The lower layers deal with the delivery acknowledge, which indicates only that the message is delivered to the Service Centre. SM-TP does not deal with the communication beyond the Service Centre, and does therefore not support an automatic indication if a message has reached its final destination. The Service Centre used by Telenor Mobil offers end-to-end acknowledgement by sending independent Short Messages.

3.2.3.2 Mobile Terminated Short Messages

A short message to a GSM telephone must first be routed from the originator to a Short Message Service Centre where a transport layer message, the 'SM-TP SMS-DELIVER', will be built. Like the Mobile Originated Messages, this message is transferred via various interfaces using underlying protocols.

Before the message can reach its destination GSM telephone, the actual routing must be derived. This is done by interrogating the location registers in GSM, which are used to keep track of the subscribers, with a special message 'SEND ROUTING INFO FOR SHORT MESSAGE'. The message is answered either by a message containing the relevant routing information, with the address of the Visiting Location Register (VLR) of the addressee or by a rejection message if the subscriber is known to be inaccessible at that instant. If the addressee is inaccessible the relevant GSM location registers (VLR/HLR) are informed that a Service Centre has an undelivered message to the subscriber. When the subscriber again attaches to the GSM network, the Service Centre is alerted and delivers the message.

The ways of accessing the Service Centre are:

- From a GSM telephone supporting Mobile Originated Short Message (MO SM)
- By using dial-up modem to the Service Centre or by direct access via a public X.25 network (Datapak). Two protocols are supported:
  1. Menu protocol for terminal emulation. This only requires a PC with a standard communication program
  2. UCP 'machine protocol' (UCP = Universal Computer Protocol), for special applications
- From a telephone with DTMF signalling. The user is guided through a voice menu. In principle, only numerical information dialled from the telephone can be put in the Short Message; alternatively in combination with preprogrammed alphanumerical texts selected from the menu.

4 Mobitex

4.1 General

The development of the Mobitex system started in the early eighties by Swedish Telecom. Today, new functionality is still evolving, but the development is handed over to the Ericsson company. The specification of interfaces to user equipment is open for all manufacturers and is handled by the Mobitex Operators Association (MOA), to which Telenor Mobil is a member.

Telenor Mobil is operating a second generation Mobitex network. The first system of this kind was opened in USA in 1991. In Norway a service based on the first generation system opened in 1989 and was upgraded to second generation in late 1993. Today, Mobitex is covering the four most populated districts of Norway.

The typical use of Mobitex is for communication between a host computer and fleet of remote, mobile terminals. The system is optimised for transferring small amounts of data traffic of 'bursty' nature (i.e. the data packets from each terminal is spread arbitrarily in time).

Some performance characteristics of the system:

- Powerful error correction protocols are used, implying a very small possibility of receiving erroneous messages
- Transit delay through the network varies with traffic load but is normally in the magnitude of 2 – 3 seconds
- Small blocking rate – if traffic congestion occurs in the network, the transit time is increased
- Net bit rate from mobile terminals is normally 1200 – 4800 bit/s depending on the application used and the traffic density of the network.

Some typical areas of use of Mobitex are:

- fleet management – dispatch
- communication with portable point of sale terminals
- telemetry – reading measuring values from and superintending remote, stationary equipment.

4.2 Services and facilities

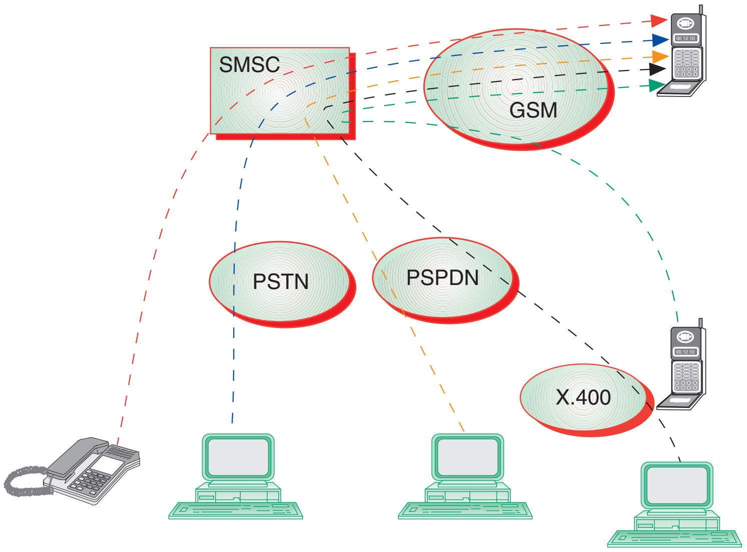
4.2.1 Message transfer

Mobile terminal equipment is connected via radio modems to the system via the Mobitex air interface, also known as 'ROSI'. The standard connection for fixed terminals is via public X-25 networks (e.g. Datapak), but other options are available. The main service of Mobitex is conveyance of packets ('datagrams') between terminals connected to Mobitex.

The packet length depends on the amount of user data to be put into the packet. The maximum amount of user data to be transmitted in one packet is 512 octets.

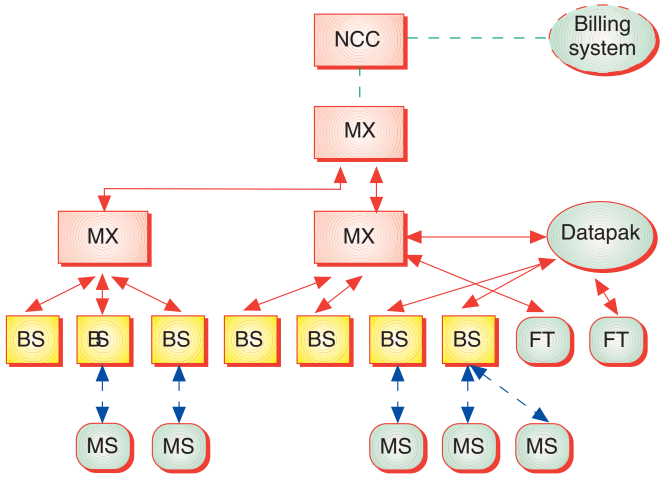
[3] *Telenor Mobil is operating a Service Centre which is manufactured by the Dutch CMG company.*

A special user packet is the ‘Status Message’. This is an empty packet with one octet of user data in the packet header. A status message then has a value between 0 and 255. These values may be used for pre-defined messages.



SMS	:	Short Message Service Centre
PSPDN	:	Public Switched Packet Data Network
PSTN	:	Public Switched Telephone Network
X.400	:	CCITT Standard for electronic mail

Figure 2 Ways of accessing the Short Message Service Centre



MS	Mobile Station	Datapak	Public X.25 network
MX	Mobitex Exchange	FT	Fixed Terminal (Host Computer)
BS	radio Base Station	NCC	Network Control Centre

- - - Connection with no user traffic

◀▶ Fixed connection carrying user traffic

◀ ▶ Radio connection carrying user traffic

Figure 3 Mobitex network hierarchy

4.2.2 Group messages

Group messages will be broadcast to all Mobitex terminals which are members of the same group. Group messages are not acknowledged by the receiver.

4.2.3 Mailbox

If the Mailbox Flag in a message header is raised, the message will be stored in the network mailbox if the addressee is temporarily unreachable, e.g. out of radio coverage. The message will then be delivered when the addressee again reaches contact with the network.

4.2.4 Positive acknowledgement

In Mobitex negative acknowledgement is standard on the network level, which implies that the message originator is notified only when the system is unable to deliver a message. For some applications this may be a disadvantage. Therefore, it is possible to be notified of delivered messages by subscribing to positive acknowledgement.

4.2.5 Closed user group

Terminals may be members of a closed user group. User group members will only be able to communicate with other members of their group.

4.2.6 Personal subscription

With a personal subscription it is possible to log on to an arbitrary Mobitex terminal. Several personal subscriptions may be logged on to the same terminal.

4.3 Network structure

The Mobitex network comprises base stations exchanges and a network control centre (Figure 2). The radio modems are connected via radio channels to the base stations. The host computers are connected to exchanges via leased line connections or public X.25 networks (Datapak).

4.3.1 Main components of the network

4.3.1.1 Base stations

The radio base stations constitute end nodes for the radio modems. They are also switching points for traffic between radio modems under the same base station. A base station is equipped with one or two radio channels depending on the traffic demand at the specific station. If the traffic over a certain base station exceeds the capacity of two channels, the geographical coverage area of that base station has to be split between more base stations, either by establishing new stations or by handing over more traffic to base stations with overlapping coverage, by adjusting network parameters.

4.3.1.2 Exchanges

Exchanges may be established at several hierarchy levels. Exchanges at the lowest level constitute as end nodes for host computers which are linked to them. They are the switching points for traffic between base stations and host computers. Higher level exchanges route traffic between lower level exchanges.

4.3.1.3 Network control centre (NCC)

NCC does not take part in the actual traffic handling. It includes an operation and maintenance function and a subscription information handler.

The operation and maintenance function is used for collecting central alarms and operating statistics, test function initiation, setting of operating parameters and program loading of all network nodes.

The subscription information is entered at the NCC and then sent to the exchanges. Charging information is collected from the network by the NCC and the necessary basis for accounting is transferred to the billing system.

4.3.2 Subscription routing information kept by the network

In order for the network to perform the traffic routing the nodes involved need to be updated with subscription specific information. This information is stored and updated by the network either as static or as dynamic information, dependent on the expected updating frequency.

Examples of static information:

- Subscription number and type
- Group numbers of which the subscriber is a member
- Services subscribed to.

Examples of dynamic information:

- Information about which base station to be used by a certain radio modem. This information is updated by the radio modem when moving from one base station to another.
- Sequence number for protection against receiving copies of the same message on the radio path
- Active/Inactive status, i.e. whether a radio modem or host computer is in contact with the network or not, for reduction of overhead traffic in the network. This information is updated by the radio modem (e.g. when powering on or off) or the network (e.g. when it fails in a message delivery).

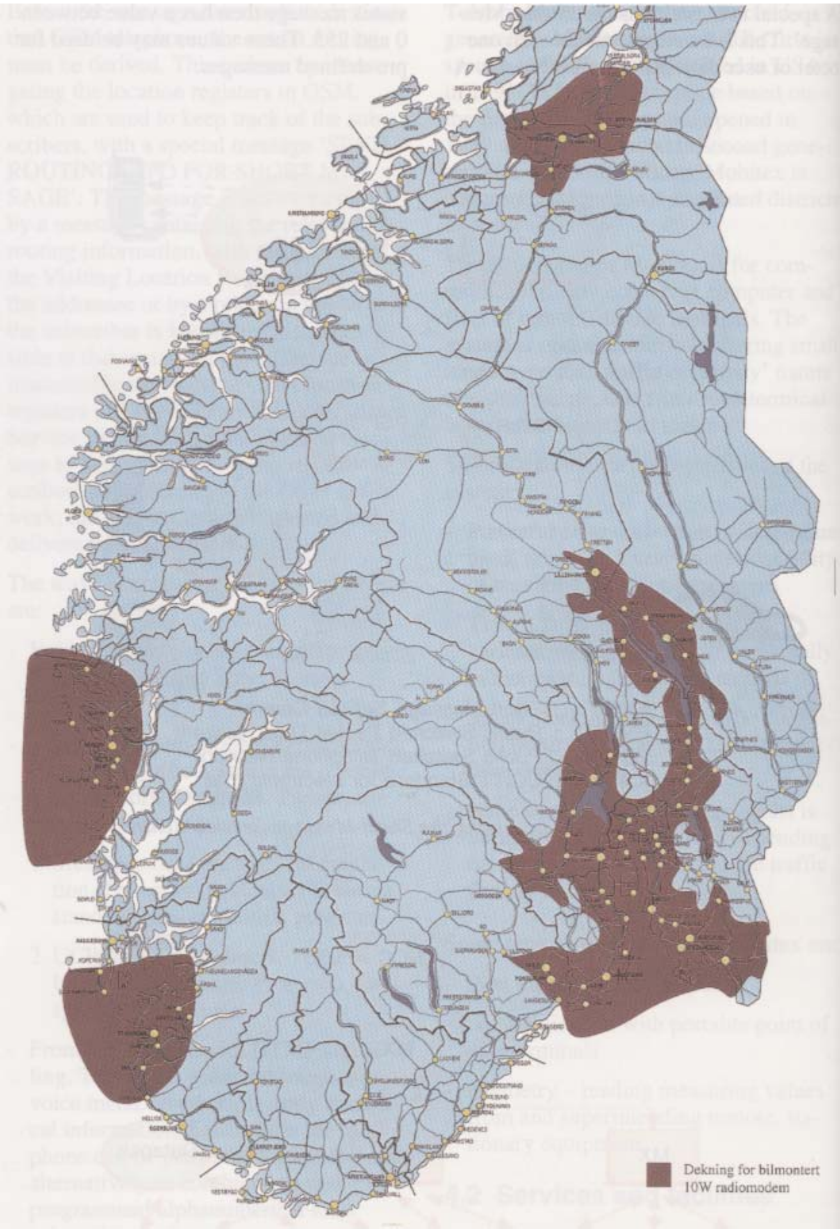


Figure 4 Map of radio coverage area

4.3.3 Radio coverage

Mobitex is covering the four most populated areas of Norway. At present 27 base stations are operating.

The radio coverage area is shown in Figure 4.

The base station sites used by Telenor Mobil were planned by help of a coverage prediction tool from Philips. This tool uses digitised geographical maps together with a radio propagation model, making is possible to do coverage simulations of several potential base station sites and antenna configurations in a short time.

Mobitex is transmitting datagrams which, if not received, are repeated. Local holes without radio coverage are often not noticed by a Mobitex user passing through, while similar holes in GSM or NMT would give serious disturbances to a voice session. Also the 400 MHz radio frequency band, used for Mobitex, is better than 900 MHz, which is used for NMT-900 and GSM, for transmission paths over long distances and has better intrusion ability into buildings. Thus, the number of base stations needed for coverage in a given area is significantly less than what is necessary for NMT and GSM.

4.4 Protocols

4.4.1 Layer 3 – the network layer

At the network layer the route selection is made and the end units are addressed, so that user data is guided correctly through the network. This part is called the Mobitex network layer and it is where all the user data are arranged as packets or MPAKs. There are basically two kinds of MPAKs:

1 MPAKs for transfer of user data – user information (0 – 512 octets of user data)

2 MPAKs for interaction with the network – system information.

The AT command set is an industry standard, widely used in North America, for circuit switched asynchronous communication.

X.28/X.3 is an asynchronous character communication / PAD standard in widespread use. X.28/X.3 is defined in CCITT's recommendations and has been specially designed for the X.25 network.

The MPAK header contains the Mobitex addresses of the sender and the addressee as well as information about packet type and the state of the packet. The normal size of the header is 12 octets. The MPAK may be a 'header only' packet or it may contain user data of variable length up to 512 octets. The user data are organised in blocks of 18 octets.

4.4.2 Layer 2 – the link layer

The link layer ensures error free transmission between the two closest interacting nodes. The separate links use different layer 2 protocols depending on the nature of that link.

4.4.3 Layer 1 – the physical layer

The physical layer deals with electrical and mechanical aspects of signalling. The separate links use different layer 1 protocols depending on the nature of that link.

4.4.4 Higher level protocols (HP)

Many applications can be built directly on the standard services used in Mobitex by using the three lower network dependent layers. Other applications demand a higher level of services. To handle this the MPAK header (layer 3) has a one octet field, 'HP-protocol', which refers to higher level protocols if carried by the MPAK. Some of the HP numbers are reserved for standardised Mobitex HP protocols.

A few higher level protocols for Mobitex are standardised, amongst them the frequently used 'MTP/1' (Mobitex Transport Protocol). MTP/1 is optimised for use in Mobitex, for particular demands in packet switched radio networks as well as for minimising the overhead. MTP/1 provides:

- Reception of data packets in the same order as received
- Division of large amounts of data into MPAKs (max. 512 bytes) for transmission. Reconstruction of the original

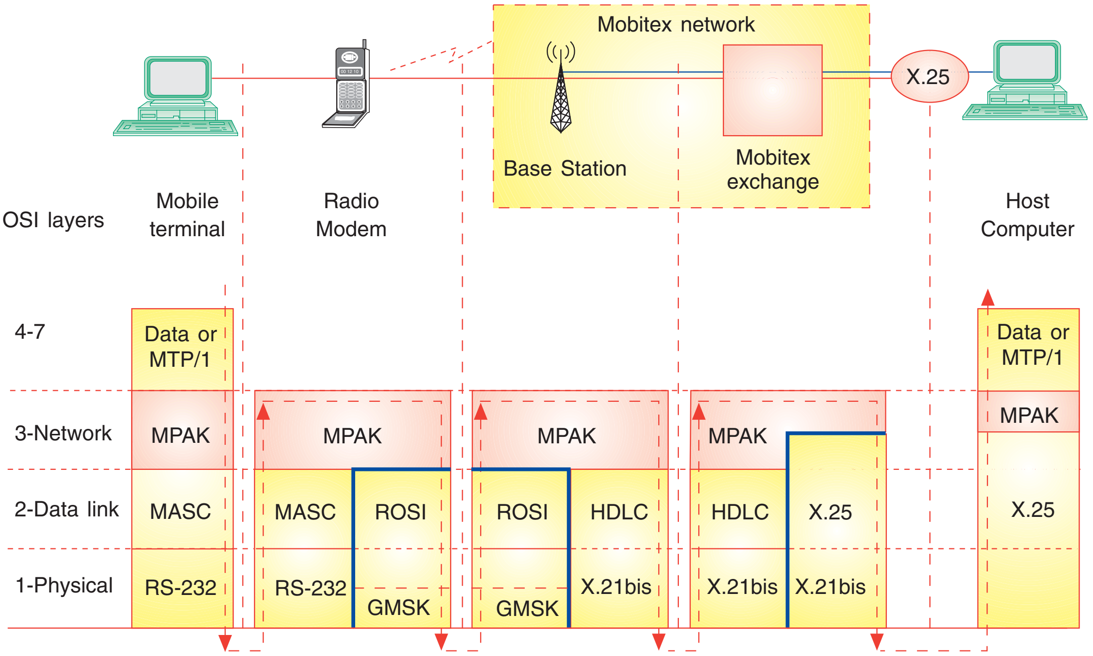


Figure 5 Mobitex protocol layers on different links

data contained in multiple packets at the receiver end

- Positive acknowledgement
- Establishing a logical connection (virtual circuit) between two user terminals.

4.4.5 Protocols on different links

4.4.5.1 Mobile data unit – Radio modem

Transfer of user data will be achieved by exchange of MPAKs on the network layer. On the physical layer RS-232 is normally used.

A special link layer protocol, MASC (Mobitex ASynchronous Communication), or machine interface, has been developed for the mobile side to withstand high levels of interference which is particularly important in vehicle use. MASC also supports functions for controlling the radio interface, such as Radio Modem ON/OFF and Network Contact OK/LOST.

From the Mobitex system point of view the MASC interface is the preferred one. However, in order to ease the connection of some existing applications and products to Mobitex, there are radio modems available with AT or X28/X3 interface instead of MASC.

4.4.5.2 Radio Modem – Base station

The Mobitex air interface, which is also known as the ROSI (RadiO SIgnalling) protocol, is used for communication between radio modems and base stations. The protocol is specifically designed to handle the unreliable radio environment in a frequency efficient way while providing the characteristics of a packet switched network.

ROSI uses a GMSK0.3 (GMSK = Gaussian Minimum Shift Keying) modulation scheme in duplex 12,5 kHz radio channels with block interleaving, CRC error detection (Cyclic Redundancy Check), a Hamming code for error correction and selective ARQ (automatic repeat request). A reservation slotted ALOHA type scheme is used for channel access that provides a good balance between efficiency and responsiveness.

4.4.5.3 Base station – Mobitex exchange

Communication between internal Mobitex network nodes may be established by use of leased lines or the X.25 network. Dial-up telephone lines may be used as backup.

4.4.5.4 Mobitex exchange – Host computer

This is the link between the Mobitex exchange and the host computer which contains the communication software of

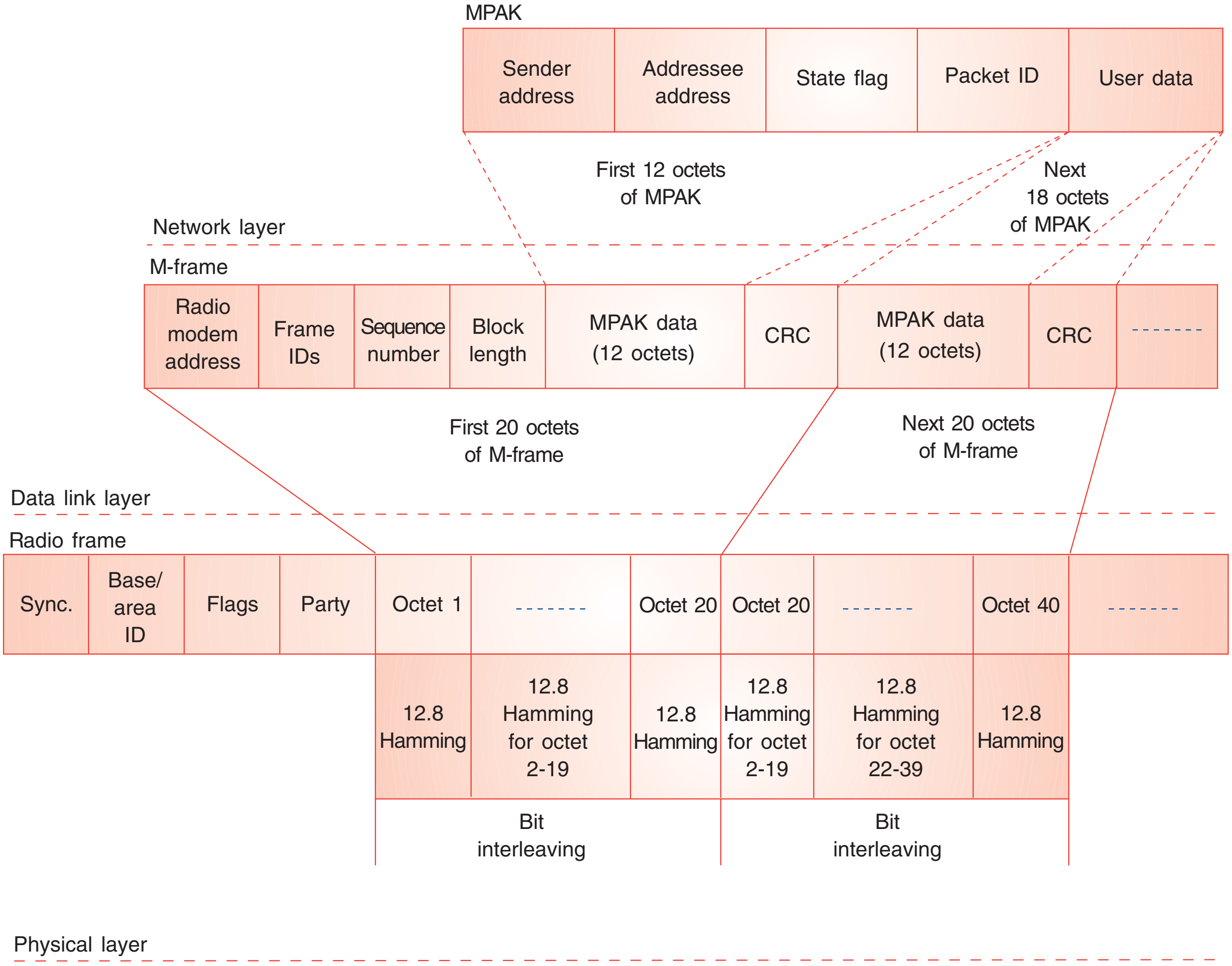


Figure 6 Mobitex ROSI frame structures

the user. This link may be established in several ways:

- Transporting MPAKs as user data via the X.25 network
- Via the X.25 network to a gateway to the Mobitex network. The Mobitex X.25 gateway transforms the X-25 user data (layer 3) to MPAKs for further transport through Mobitex and vice versa, thus providing connectivity with standard X.25 hosts
- Transporting MPAKs via leased lines with HDLC or MASC on the link layer
- Transporting MPAKs via a Mobitex radio modem ('mobile host computer'). This solution is only recommended for trials.

The two X.25 alternatives are the recommended ways of host connection to Telenor Mobil's Mobitex network.

References

1 Mouly, M, Pautet, M-B. *The GSM system for mobile communications.* Palaiseau, published by the authors, 1992, ISBN 2-9507190-0-7.

2 Bjugan, V. *GSM : det globale systemet for mobilkommunikasjon.* Trondheim, Tapir, 1995. ISBN 82-519-1190-7. (In Norwegian.)

3 ETSI. *Technical realisation of the short message service (SMS) point-to-point, rev. 4.10.0.* 1994. (ETS 300 536.)

4 *Mobitex operators association : Mobitex interface specification rev. R3A.* Ericsson.

5 *Mobitex connectivity catalogue.* Ericsson, 1994.

6 Kahn, M, Kilpatrick, J. Mobitex and mobile data standards. *IEEE communications magazine,* 1995.

7 Levesque, A H, Pahlavan, K. Wireless data communications. *Proceedings of the IEEE,* 82(9), 1994.

8 Alanka, T et al. *Measured performance of data transmission over cellular network*. Department of computer science, University of Helsinki, 1994. (Report C-1994-53)

Standards for wide-area paging – the situation in Norway and Europe

BY LARS A BØRVE AND PER HJALMAR LEHNE

In the shade of mobile telephony, paging, also a mobile service, is popular both as an add-on to mobile telephones and as a stand-alone service. Alphanumeric paging, the possibility of sending text messages to a small pocket receiver, has opened new markets for paging. Also the domestic market seems to have "discovered" paging.

This article explains the two leading standards for public, or wide-area paging, POCSAG and ERMES, together with a presentation of the Norwegian situation. Some remarks are also done about the future for paging.

1 Introduction

Telenor's subsidiary Telenor Mobil AS, started its first tone-only and numeric paging service, "PS-tall", in December 1984. The system was delivered by NEC, Japan. A combined tone-only, numeric and alphanumeric paging service, "PS-tekst", was launched in June 1991 delivered by the Finnish company Tecnomen. Today, Norway has got the highest penetration (approximately 3 %) of paging in Europe.

1.1 The first international paging standard

The work with the radio code on which the two Norwegian paging services are based, started in 1976. An international group of engineers met to explore the possibility of agreeing on a mutually acceptable code for wide area paging. The first meeting was held in London and chaired by the then British Post Office. The subsequent success of the meetings that followed resulted in the POCSAG code (The Post Office Code Standardisation Advisory Group) [1]. Official international recognition of this code was given in 1981 when it was accepted by the CCIR as the recommended Radiopaging Code No. 1 (RPC 1, Rec. 584) [2]. More than 10 million POCSAG pagers have been sold, and almost all paging manufactures offer infrastructure and terminals conforming to the standard. POCSAG enables the use of simulcast transmission with data speeds of 512 bits per second (bit/s), 1200 bit/s and 2400 bit/s. The two lowest speeds have been used for a while, while the use of 2400 bit/s has started recently. Simulcast transmission means that the page is sent from all transmitters at the same time (bit level).

1.2 Next step: pan-European paging

In spite of the commercial success for the POCSAG code, it only offers national coverage. Aroused from the spirit of the GSM work, the need for a paging system also offering services at an international level was recognised. In April 1986 the Telecommunication Commission of CEPT established the R35 Subworking Group with the task of defining a new pan-European radiopaging system capable of allowing roaming at European level and guaranteeing for compatibility of paging receivers throughout Europe. Later, with the reorganisation of the radio activities of CEPT, R35 SWG changed its name to RES4. RES4 used three specialised 'ad hoc' working groups to speed up the work: Service and Facilities (SF), Radio Sub-System (RSS) and Network Aspect (NA). From November 1988, RES4 became a Sub-Technical Committee (STC) under the new ETSI structure and soon reorganised to be the PS (Paging Systems) Technical Committee (TC), and the three 'ad hoc' working groups became Sub-Technical Committees PS1, PS2 and PS3, respectively.

The working name for the system was European Radio Message System, abbreviated ERMES, a name which later was adopted as the official name of the system. The official logo is shown in Figure 1.

The work of these STCs was basically finished by the end of 1990, and ERMES was adopted as a European Telecommunication Standard (ETS) in 1992 [4] – [10].

The International Telecommunication Union (ITU) approved ERMES as the world's first recommended standard for paging systems intended for international use in September 1994 (ITU-R M.539-3 [11]).

Figure 1 The official ERMES logo

2 Norwegian paging networks and services

In Norway two different paging networks are in operation – one tone-only and numeric service and one alphanumeric service. The old tone-only/numeric network operates on two frequencies, 148.050 MHz and 148.100 MHz, sharing the airtime, using two time slots. This system covers approximately 85 % of the Norwegian population. The combined tone-only, numeric and alphanumeric paging network operates on the frequency 169.800 MHz and covers approximately 80 % of the population. Both systems are based on the POCSAG radio code. The coverage as per January 1995 is shown in Figure 2.

2.1 The network architecture

The infrastructure for the two paging networks are similar to each other. Both systems consist of a paging terminal (PT), district controllers (DCs), base station controllers and radio transmitters. The alphanumeric system contains a network controller (NC) as well.

2.1.1 The Paging Terminal (PT)

The main task for the paging terminal (PT) is to communicate with the calling party, i.e. see to it that a page accepted by the PT is sent to the right pager in the subscriber's paging area(s). The page can be generated either in the numeric interface, using DTMF [12], or one of the alphanumeric interfaces (X.25 or dial-up modems).

When sending a page, the sender would have to transmit the page according to a certain protocol. Some of these may be open, international standards or proprietary protocols. The protocol describes the dialogue between the sender and the PT and all the situations that can occur while connected to the PT.

The general input procedure is:

1 Connection opened

2 Subscriber number entered

3 The subscriber number is compared with the information stored in the subscriber database

4 If valid, the PT accepts the subscriber number

5 Entering of the message

6 The message is terminated

7 PT checks the message (length, etc.)

8 PT accepts or rejects the message

9 Connection broken.

When accepted, the page will be stored in the correct buffer, dependent on which one of the eight frames the pager belongs to, the transmission speed to be used, the paging area(s) it belongs to, etc. The content of these buffers are forwarded to the network controller which will control the transmission of the pages (this operation is done by the PT in the NEC system).

The PT is connected to VDUs that are used during configuration and reconfiguration of the PT's general parameters. Such parameters can be the time between transmission of pages from PT to network controller, how long the PT shall store a page if the page cannot be forwarded to the transmission network if it is down) etc. The PT is connected to the subscriber and billing system, and all relevant information about the subscription is fed through this interface. Such information can be radio identity code (RIC), baud rate, paging area(s), subscriber number, pager type, etc.

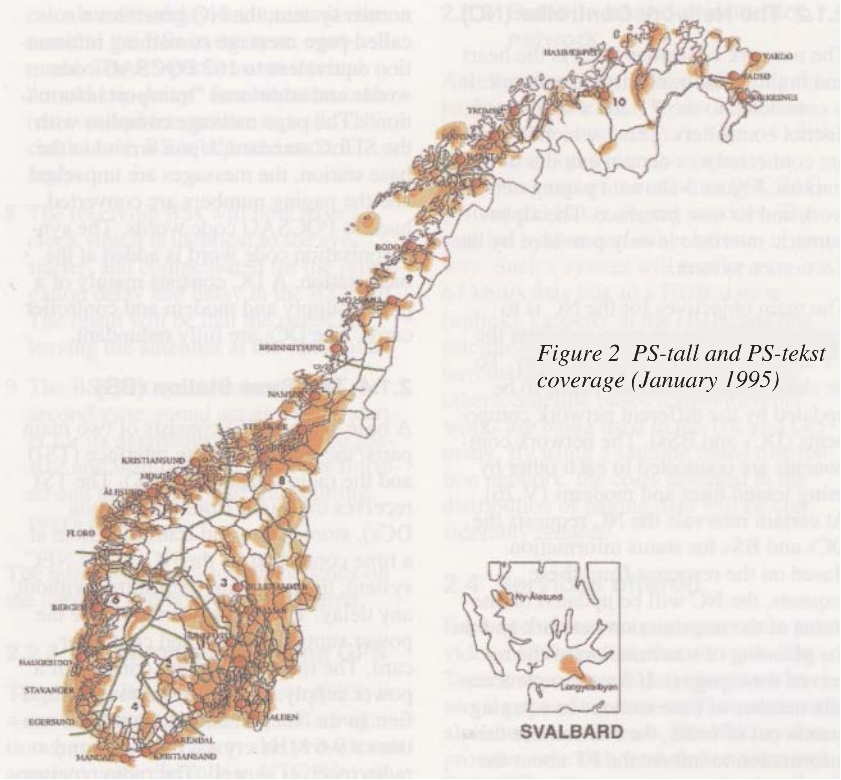


Figure 2 PS-tall and PS-tekst coverage (January 1995)

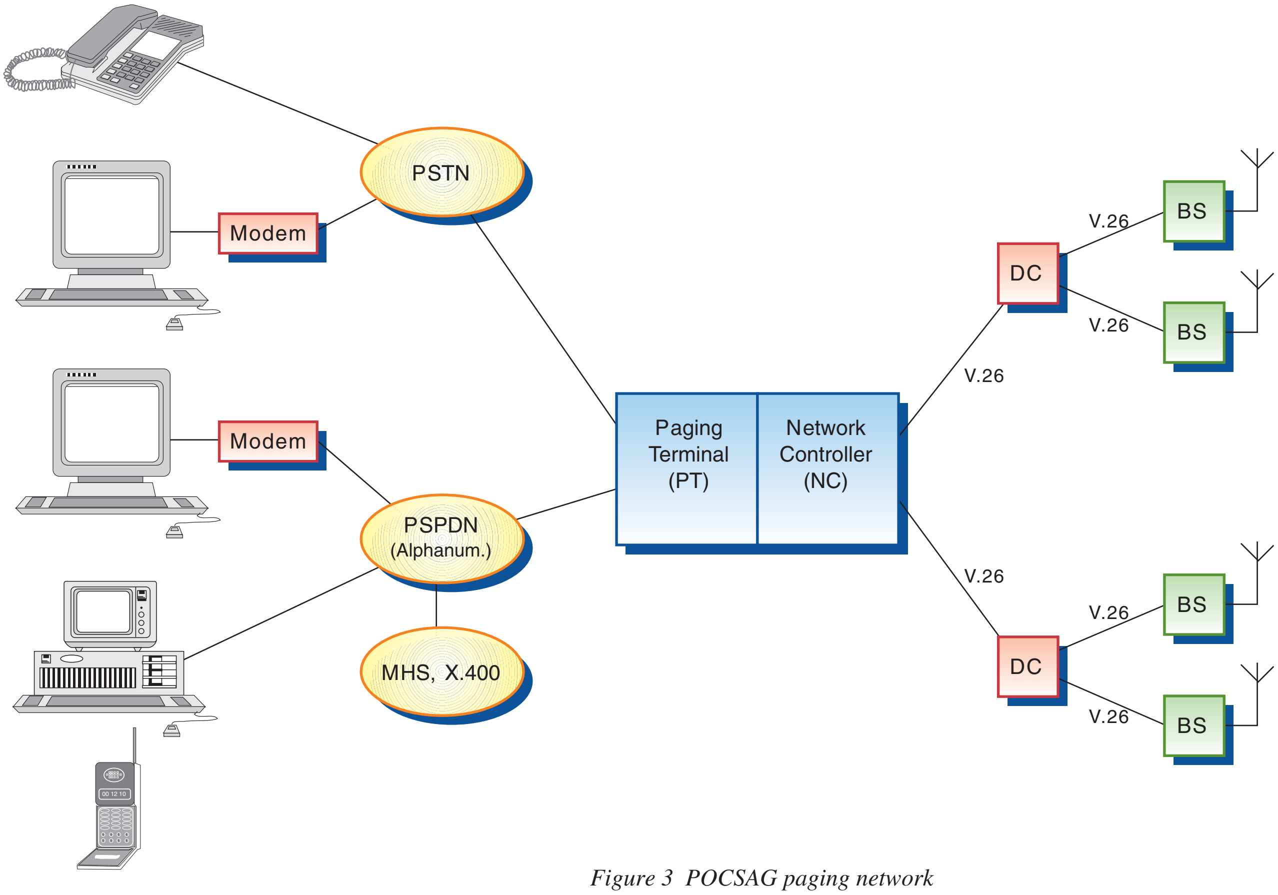


Figure 3 POCSAG paging network

2.1.2 The Network Controller (NC)

The network controller (NC) is the heart and brain in the transmission network. It is connected to the PT and a number of district controllers (DCs), which in turn are connected to a certain amount of base stations. Figure 3 shows a paging network and its user interface. The alphanumeric interface is only provided by the Tecnomen system.

The main objectives for the NC is to ensure that the pages received from the PT are transmitted in a correct way. To be able to do this, the NC needs to be updated by the different network components (DCs and BSs). The network components are connected to each other by using leased lines and modems (V.26). At certain intervals the NC requests the DCs and BSs for status information. Based on the response from these requests, the NC will be updated on the status of the transmission network and do the planning of transmission of the received data (pages). If for example a certain number of base stations in a paging area is out of order, the NC can use this information to inform the PT about the situation in the paging area. The PT will then inform the calling party about the situation in the area in which the page is to be sent.

The NC is connected to an alarm printer and a VDU. The VDU is used to configure and change the parameter settings for the NC and the transmission network. The NC (PT for the NEC system) can be connected to maximum 14 DCs. A paging area may consist of one or more DCs.

The NC is fully redundant, i.e. if a faulty situation occurs in the operating part of the NC, an immediate switch-over to the slave side will take place, without losing any information (hot stand-by).

2.1.3 The District Controller (DC)

The district controllers (DCs) operate as nodes (line concentrators) between the NC and the base stations. The DCs are used to avoid many long and expensive leased lines between the NC and the BSs. The DCs' main task is to forward the pages from the NC to the BSs for transmission. All base stations under a specific DC's control transmit the same page, at the same time on the same frequency.

For the NEC system, the data stream from the PT is converted to POCSAG format in the DC, and the synchronisation code word is added here. In the Tecnomen system, the NC generates a so-called page message containing information equivalent to 162 POCSAG code words and additional "transport information". The page message complies with the SDLC standard. Upon arrival at the base station, the messages are unpacked and the paging numbers are converted back to POCSAG code words. The synchronisation code word is added at the base station. A DC consists mainly of a power supply and modem and controller cards. The DCs are fully redundant.

2.1.4 The Base Station (BS)

A base station (BS) consists of two main parts; the transmitter site interface (TSI) and the radio transmitter (TX). The TSI receives the pages from the NC (via DCs), stores them and transmits them at a time controlled by the NC. In the NEC system, the pages are transmitted without any delay. The TSI's main parts are the power supply, modem and controller card. The radio transmitter consists of a power supply, exciter and a power amplifier. In the Tecnomen system the BS contains a 9.6 MHz crystal oscillator and a radio receiver as well. The radio receiver is used during the radio synchronisation of the network. The output power from the transmitters are in both cases 100 Watts.

The TSI performs self diagnostic functions and reports back to NC (via DC) about any faulty situations, which in turn will generate an alarm. The modems used in the BSs are 512 bit/s (NEC) and 2400 bit/s (Tecnomen).

The BSs are non-redundant, i.e. if a major fault occurs, the complete BS will go out of operation.

2.2 Network synchronisation

When base stations in an overlapping area are sending the same page, the subscriber receiver may receive the same message from several base stations. Thus, the subscriber receiver may receive the message erroneously as a sum of several separate messages. The correctness of the received message is dependent on how much out of phase the bit(s) are upon reception. To minimise the phase difference between neighbouring base stations, different synchronisation techniques can be applied.

Investigations have shown that the worst case reception of a message, sent from two transmitters, occurs when these two conditions are fulfilled:

1 The receiver experiences a difference in field strength between the two radio signals which is less than 3 dB.

2 The two signals have a minimum phase difference of at least 1/4 of a bit (239 µs at 512 bit/s,100 µs at 1200 bit/s).

Condition 1 is fulfilled in practice when a receiver is located within the overlapping area of the two transmitters. The worst case overlapping area is usually estimated to be approximately 3 km.

In order to meet the requirements stated in 2, the transmitters should be synchronised with each other. Different techniques can be applied to meet these requirements, and some alternatives are mentioned below.

2.2.1 Line synchronisation

To ensure a synchronised transmission of pages, the NEC system uses line synchronisation. Since we have an immediate transmission of pages received at the NEC base station, the transmission will not be delayed in the BS. To avoid > 239 µs phase difference (512 bit/s) between neighbouring base stations, and to "synchronise" the BSs, the following is done in the NEC network:

1 The line delays between the PT and each of the DCs, and between the DCs and BSs, are measured.

2 All modems (transmission connection card) are adjusted to compensate for the line delay.

3 All pages will now be received at the BSs at the same time, hence transmitted by the BS at the same time.

The drawback of this method is the maintainability of the network. If e.g. a leased line between a DC and a BS is given an alternative routing, the line delay would need to be measured and a retuning must be performed. Different switching times (and delays) in the telephone network will seriously affect the result of this method.

2.2.2 Radio synchronisation

In the Tecnomen system, radio synchronisation is used. The main idea behind this method is to synchronise the clocks in the transmitters so that a message will leave the BS antennas at the same time, and so that there are no significant phase differences between signals transmitted from two BSs with overlapping coverage areas.

The radio synchronisation is executed using the same frequency as the actual transmission of paging messages. The use of the same frequency will reduce the capacity of the paging system. However, if the time between each synchronisation round is made as long as possible, and the synchronisation is performed as quickly as possible, the drawback is minimised.

During the radio synchronisation the following actions will be performed:

1 The NC is given information about the whole network, i.e. the position of each BS, the distance between neighbouring BSs, etc.

2 This information and information about the status of the network, is used to calculate the synchronisation plan, determine which BS(s) that shall act as the so-called sync. starter, etc.

3 The radio synchronisation starts with the NC sending out (using the leased lines) its internal clock information to the BSs (via DCs). Due to different distances and line delays, this time information is received at the BSs at different times and with varying degree of error. Now the BSs will have approximately the same time information.

4 The NC sends out a synchronisation command to the BS chosen as sync. starter. At the same time the BSs that are in radio contact with the sync. starter, will receive a command (from NC) to get ready for reception of the sync. message. The sync. message contains information about when the sync. message will be transmitted from the sync. starter, information about the propagation delay between sync. starter and the BS in question, etc.

5 The receiving BSs will now make a switch-over from acting as transmitters, to acting as receivers, and listen for the sync. message.

6 When the sync. message is detected, the receiving BSs will compare the time information given in the sync. command and the received sync. message. This information will be used to calibrate the BSs' internal clock.

7 The next task is to compensate for the propagation delay between the sync. starter and the receiving BSs. The NC has got the necessary information to calculate this delay, and transmits the result to the BSs in the sync. command. The BS performs an internal "round-trip delay" measurement in the receiver and transmitter as well. The clock in the BS is adjusted accordingly.

8 The receiving BSs will now have a clock which is identical to the sync. starter, and compensated for the propagation delay and delay in the BS itself. The result will be that the pages are leaving the antennas at the same time.

9 The BSs just synchronised can in the second sync. round act as a sync. starter for its neighbouring BSs, or go passive and wait until the sync. has finished and the BS can start transmitting pages again.

The number of sync. rounds depends on the network and how it is configured.

2.2.3 Synchronisation using GPS

The Global Positioning System (GPS) is a satellite based positioning and navigation system. This system also provides an accurate time reference to UTC (Universal Time Coordinated). The advantage using GPS is that all base stations can be synchronised individually, 24 hours a day. This time information can be used in different ways to tune the clock in the BS so that the pages will leave the antenna at approximately the same time.

2.3 The distribution network

In order to minimise the transmission costs, different ways of distributing the paging information must be considered. Traditionally, leased lines with modems have been used in the distribution network. However, the use of leased lines is expensive for the paging operators, and cheaper and equally stable solutions are available.

2.3.1 Sharing capacity with cellular systems

On most paging sites, there is a cellular base station as well. By using multiplexers it is possible to share the network capacity to this site with the cellular network. If for example there is a 2 Mbit/s link to a BS, one (64 kbit/s) of the 30 available channels (time slots) can be used for paging and the rest for the cellular or other services. Telenor Mobil will gradually move from the use of traditional leased lines to the solution mentioned above.

2.3.2 Satellite based distribution network

An even cheaper way of distributing the paging information is to use a satellite based distribution network. One 64 kbit/s channel will give enough capacity to cover the need of the existing POCSAG system and future paging services. The satellite system can either be one- or two-way. Such a system will require a 64 kbit/s data link to a HUB station (uplink), capacity in the HUB, and one satellite terminal for each of the BSs. The terminals used can be either VSAT or others. Satellite based distribution networks are being used in the UK and Germany. By using a satellite based distribution network, the costs involved in the distribution of paging data will be considerably reduced.

2.4 Services offered

Each one of the two paging services provides different subscriber services. The Tecnomen system is an alphanumeric service providing tone-only, numeric and alphanumeric services. The NEC system provides tone-only and numeric services. The NEC system covers approximately 85 % of the population, while the figure for the Tecnomen system is approximately 80 %. The NEC system is divided into 10 paging areas, and a subscriber can subscribe to either one, two, or three paging areas, or nation-wide. In the Tecnomen system there is only one paging area (nation-wide).

2.4.1 Tone-only paging

In the POCSAG standard, one RIC can be used to initiate 4 different beeps or alert types. Each alert type is registered as a separate pager number. Depending on which one of these pager numbers are called, a unique beep is generated in the pager. A call is generated from DTMF telephones. A typical use of tone-only paging might be that one of the subscriber numbers are used by your company, another for domestic use, a third is used by a friend, etc. Another use of tone-only paging, is to let your pager alert you whenever someone has called your voice-mailbox. The NEC and Tecnomen systems provide the same tone-only service.

2.4.2 Numeric paging

The most popular service is numeric paging. This service provides you with the ability to send 10 characters (NEC) or 15 characters (Tecnomen). A typical use for

/ 1 A B C	SP 2 D E F	U 3 G H I
- 4 J K L	- 5 M N O	% 6 P Q R
7 S T U	8 V W X	9 Y Z
* Esc Ai Num	0 SP ?	#

Figure 4 Alphanumeric keyboard layout according to CEPT T/CS

this service is to send (using DTMF telephones) your telephone number to a numeric paging subscriber, so that he/she can call you back. A numeric pager can also be used in conjunction with the voice-mail system (alerting).

2.4.3 Alphanumeric paging

Alphanumeric paging is only provided by the Tecnomen system. Messages containing up to 128 alphanumeric characters may be sent to an alphanumeric pager. An alphanumeric message may be sent from a number of different inputs, which will be discussed below. The PT provides two different alphanumeric protocols, Tecnomen alphanumeric protocol and Telocators alphanumeric protocol (TAP), each of which has got its own modem pool with a given access number. In addition to this, the Tecnomen protocol is implemented for X.25.

2.4.3.1 Bureau service

By calling the telephone number **96 89 00 10**, you will be connected to a bureau service. They will ask to which pager you want to send a message, and the content of the message. When the message has been recorded, it will be sent by the operator.

2.4.3.2 Predefined alphanumeric pages

Another way of sending alphanumeric messages is the predefined message service. From the VDU in the PT, the operator can define, and edit, 100 predefined messages, each of which is numbered from 00 to 99. To send a predefined message you have to call the telephone number **96 89 00 20**, key in the subscriber number (the last five digits), and the two-digit combination which indicates the predefined message you want to send. It is possible to send just one message at a time, i.e. different predefined messages cannot be combined and sent as one single page.

2.4.3.3 Alphanumeric DTMF interface

A DTMF-telephone can also be used to send alphanumeric messages. It is a quite complex and tedious way of sending a message. The concept is based on the use of the DTMF-tones. The user makes a call to the telephone number **96 89 00 40**, the subscriber number is entered, and the sender is asked to enter the message. The procedure starts by entering the sequence **#. Thereafter, the message is entered according to CEPT T/CS. The message is terminated with the #-key. Keying the sequence ***# takes you back to the numeric mode. Figure 4 shows the coding of DTMF characters according to CEPT T/CS.

2.4.3.4 Modem interface

In the Tecnomen system we have two different alphanumeric communication protocols, each of which is connected to a modem pool. The access number for Tecnomen alphanumeric protocol is **96 89 00 30**, and for the Telocators alphanumeric protocol (TAP) the access number is **96 89 00 50**. An in-depth discussion of these protocols is beyond the scope of this article, but a general comment is that the Tecnomen protocol is interactive (man-to-machine), while TAP is a machine-to-machine protocol. A number of different software firms have produced applications based on these protocols. These applications run on Apple Macintosh, PCs, PC-networks, UNIX-based platforms, MS-Windows etc. Most of these applications have a similar user interface; local subscriber lists, group lists, log files, communication configuration, help functions, etc.

2.4.3.5 Public Switched Packet Data Network – PSPDN

The alphanumeric, Tecnomen PT, also contains an X.25 interface. This interface supports the Tecnomen alphanumeric protocol only. If a PC or a terminal is connected to the PSPDN (X.25), it is quite easy to send a page. The PSPDN-access number for this service is **530307**. The pages generated in e.g. the Videotex service and the MHS service, are forwarded using this interface.

2.4.3.6 Message Handling Systems (MHS) – X.400 interface

The X.400 standard is becoming increasingly popular. Most of the electronic mail systems today are X.400 based. It is used to transfer documents, telefax, paging messages, etc. Together with TelePost Communication, which is an X.400 operator in Norway, Telenor Mobil offers a gateway between the X.400 customers and the alphanumeric paging system's X.25 interface. The gateway converts the received data from X.400-format to a suitable format which can be used for transmission of pages in the alphanumeric system. The general X.400 address to a "PS-tekst" pager is:

S = <5-digit subscriber number>;
PRMD = ps-tekst;
ADMD = telemax;
C = no;

2.5 Future services

The possibility to send and receive alphanumeric messages opens for a lot of exciting possibilities for alphanumeric paging. By implementing the same radio identity code (RIC) into many pagers, it is possible to send a message to an infinite number of pagers, using the same airtime as for sending just one message. This technique is used when e.g. foreign and stock exchange updates are sent out to a number of pagers. The subscriber can use the same pager for ordinary alphanumeric paging. If a person subscribes to the foreign and stock exchange information, his subscription can be controlled over the air. Controlled means in this case that the subscription can be open, closed or changed by sending a reprogramming message over the air. The same technique can be used to send other types of information. Some pagers may have 4 RICs, each of which can be divided into 4 subaddresses, using the four different function bit combinations, i.e. a pager can have 16 different subaddresses that can be used to receive 16 different types of information. If the pager is used as an ordinary pager as well, only 12 subaddresses will be available for news or other types of information.

Another concept based on paging technology is EMBARC (Electronic Mail Broadcast to A Roaming Computer). The idea behind this concept is to send E-mail, memos, spreadsheets, news services to a computer connected to a Newstream pager via an RS232-port. The Newstream pager can store a maximum of 32 kbytes and one message may contain a maximum of 1500 characters. The Newstream pager may in the future be replaced by a PCMCIA pager. Looking even further, trials with two-way paging is already taking place in the USA.

A huge growth is expected in the demand for tone-only and numeric paging in the domestic market. Wristwatch pagers are now available on the market (Swatch and Motorola), and will become very popular on the domestic market if the price is right.

3 The POCSAG radio code

With reference to Figure 5, a transmission starts with a preamble, followed by the synchronisation code word and 16 address and message code words.

3.1 Preamble

A transmission will start with the preamble. The preamble consists of a minimum of 576 bits with alternating zeros and ones (101010...). With a 512 bit/s transmission, the preamble will last for more than 1.0625 seconds. The preamble will provide bit synchronisation and will be active for a few milliseconds. If the pager does not "hear" the preamble, it will switch itself off for 1.0625 seconds (at 512 bit/s). This mechanism will provide the system with battery saving capabilities.

3.2 Batch structure

The preamble is followed by a number of batches. Each batch starts with the synchronisation code word, followed by 8 frames. One frame consists of two code words, i.e. one batch equals 17 code words. The frames are numbered from 0 to 7, and the whole pager population is divided into 8 groups, i.e. each pager is allocated to one of the 8 frames. If the pager's Radio Identity Code (RIC – 7 decimal digits) is decoded to 21 bit binary code, the 3 least significant bits (lsb) tell us which frame the pager belongs to. If these 3 bits are 000, the pager belongs to frame 0, if the bits are 101, the pager belongs to frame 5, etc.

3.3 Synchronisation code word

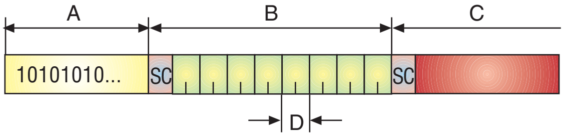
If the pager can "hear" the preamble, it will wait for the synchronisation code word to come. The synchronisation code word is a predefined bit pattern of 32 bits (Figure 6).

The synchronisation code word will provide synchronisation on the frame level (frame synch.), which is the reason why the synchronisation code word is sent at the start of the batch. When the pager is synchronised on the frame level, it will switch itself off and wait until the time has come for its own frame. Then it turns itself on again to check if one of the two code words in that frame is its address code word or not. This mechanism provides another battery saving facility. The pager will be on for just 3/17 of the time in the batch.

3.4 Address code word

Every page starts with an address code word, followed by the message code word. The address and message code words are shown in Figure 7.

Each pager has a unique address, a 7 digit decimal number (Radio Identity Code, RIC). When converted to binary code (BCD), the result is the pager's 21 bit address. Only the 18 most significant bits are transmitted on the radio path (I), the three least significant bits give the Paging Terminal (PT) information about which frame (8 possibilities) the pager "belongs" to. Such information will be used by the PT when it performs the packing of pages (to different frames) before they are forwarded in the transmission network. In addition to these 18 address bits (I), an address code word consists of a flag bit (always a binary "0"), which is sent as the msb (H), and two function bits (J). The function bits

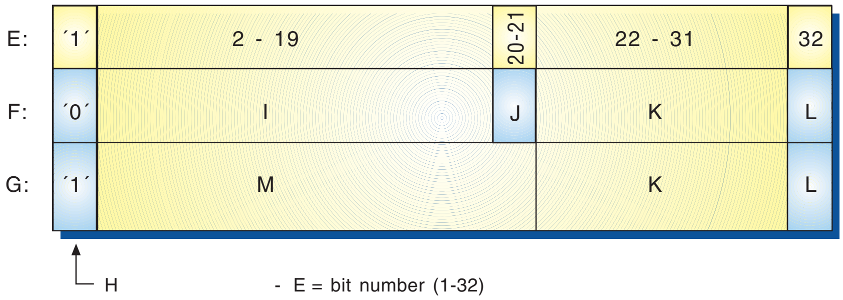


A = Preamble
B = First batch = 1 Synch. codeword and 8 frames
C = Second batch
D = One frame = 2 codeword = 2•32 bits
SC = Synchronisation codeword = 32 bits

Figure 5 POCSAG transmission scheme

Bit No.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Bit	0	1	1	1	1	1	0	0	1	1	0	1	0	0	1	0
Bit No.	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
Bit	0	0	0	1	0	1	0	1	1	1	0	1	1	0	0	0

Figure 6 POCSAG synchronisation code word



- E = bit number (1-32)
- F = address codeword
- G = message codeword
- H = flag bit (0=adress codeword, 1=message codeword)
- I = address bits (2-19)
- J = function code bits (20-21)
- K = check bits
- L = even parity bit
- M = message bits (2-21)

Figure 7 POCSAG address and message code words

Bit No.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Bit	0	1	1	1	1	0	1	0	1	0	0	0	1	0	0	1
Bit No.	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
Bit	1	1	0	0	0	0	0	1	1	0	0	1	0	1	1	1

Figure 8 POCSAG idle code word

tell the pager which subaddress of the pager the message belongs to.

The function bits are followed by 11 check bits (K). Check bits are used to enable error detection and correction. Errors may occur due to fading or noise on the radio path, while the page is being transmitted. Check bits are used in the message code word (G) as well.

The final bit (L), an even parity bit, terminates the address code word.

3.5 Message code word

The differences between the address code word and the message code word are that the flag bit (H) is always set to "1", and that the function bits are replaced by 2 more message bits. Thus, the message code word consists of 20 message bits (M). These bits are the available capacity when sending numeric and alphanumeric information in a single message code word. If a message (page) is longer than these 20 bits can provide, the message will continue into the next frame(s), and even into the next batch(es), only separated by the synchronisation code word.

Combination b3...b0	Character	Combination b3...b0	Character
0 0 0 0	0	1 0 0 0	8
0 0 0 1	1	1 0 0 1	9
0 0 1 0	2	1 0 1 0	(spare)
0 0 1 1	3	1 0 1 1	U
0 1 0 0	4	1 1 0 0	<SP>
0 1 0 1	5	1 1 0 1	-
0 1 1 0	6	1 1 1 0	]
0 1 1 1	7	1 1 1 1	[

Figure 9 POCSAG numeric character set

3.6 Idle code word

If there are no address or message code words to be transmitted, idle code words are transmitted. An idle code word looks like an address code word, but is a unique 32 bit combination, reserved for this purpose. The content of the idle code word is shown in Figure 8.

3.7 Code word generation

The error correction coding is based on the BCH (31,21) code [14]. Each code word consists of 21 information bits, corresponding to the coefficients of a polynomial with terms from X30 to X10. This polynomial is divided by the generating polynomial:

$$g(X) = X^{10} + X^9 + X^8 + X^6 + X^5 + X^3 + 1 \quad (3.1)$$

The check bits correspond to the coefficients of the terms from X^9 to X^0 in the remainder polynomial found at the completion of this division (using modulo-2). The Hamming distance for the code is 6. Using hard decision decoding any randomly displaced 5 bit errors in a code word can be detected, and any 2 bit error patterns corrected.

3.8 Message format

When sending a message (a page), 20 bits will be the available capacity in each message code word. The formats described below are considered to be the standard format.

3.8.1 Numeric message format

Numeric messages may consist of decimal numbers, spaces, hyphens (-), opening and closing brackets ([ ]), an urgency symbol (U), and one unused character (spare). These 16 characters are represented using 4 bit Binary Coded Decimal (BCD) (see Figure 9).

When sending a numeric message, the function bits are set to 00.

3.8.2 Alphanumeric message format

The alphanumeric message format is defined according to CCITT Alphabet No. 5 [3]. Each character is represented by a 7 bit combination, giving the total number of 128 different characters. When sending an alphanumeric message, the two function bits are set to 11.

3.9 Modulation

The modulation used in POCSAG is FSK (Frequency Shift Keying) – Non-return to zero (NRZ). The nominal system frequency deviation will be ± 4.5 kHz in a 25 kHz channel. The modulation rise time is 250 µs ± 25 µs (10 % – 90 % frequency transition time).

4 Second generation paging – ERMES

ERMES is different from POCSAG, both in a formal way and in complexity and service level.

While POCSAG only defines a radio code, ERMES is a true system standard covering all parts of the paging network. The main goal is the possibility of common equipment standards, and the use of a common frequency band throughout Europe. The consequences of this is the need for standardising all entities and interfaces in the paging network, not only the radio code. Standards for equipment performance had to be set and measuring methods described.

The ERMES standard ETS 300 133 comprises 7 parts:

Part 1: General Aspects [4]

Part 2: Service Aspects [5]

Part 3: Network Aspects [6]

Part 4: Air interface specification [7]

Part 5: Receiver conformance specification [8]

Part 6: Base station conformance specification [9]

Part 7: Operation and maintenance aspects [10]

Additionally, some other documents have been prepared by ETSI:

Technical Basis For Regulation, TBR 007: Receiver specifications [13]

ETSI Technical Report, ETR 050: European Radio Message System (ERMES) [14]

Combination b3...b0	Character	Combination b3...b0	Character
0 0 0 0	0	1 0 0 0	8
0 0 0 1	1	1 0 0 1	9
0 0 1 0	2	1 0 1 0	/
0 0 1 1	3	1 0 1 1	<SP>
0 1 0 0	4	1 1 0 0	U
0 1 0 1	5	1 1 0 1	-
0 1 1 0	6	1 1 1 0	.
0 1 1 1	7	1 1 1 1	%

Decimal	Char	Decimal	Char	Decimal	Char	Decimal	Char
00	@	32	<SP>	64	i	96	¿
01	£	33	!	56	A	97	a
02	$	34	"	66	B	98	b
03	¥	35	#	67	C	99	c
04	è	36		68	D	100	d
05	é	37	%	69	E	101	e
06	ù	38	&	70	F	102	f
07	ì	39	´	71	G	103	g
08	ò	40	(	72	H	104	h
09	Ç	41	)	73	I	105	i
10	<LF>	42	*	74	J	106	j
11	Ø	43	+	75	K	107	k
12	ø	44	,	76	L	108	l
13	<CR>	45	-	77	M	109	m
14	Å	46	.	78	N	110	n
15	å	47	/	79	O	111	o
16	Δ	48	0	80	P	112	p
17	<DC1>	49	1	81	Q	113	q
18	Φ	50	2	82	R	114	r
19	Γ	51	3	83	S	115	s
20	Λ	52	4	84	T	116	t
21	Ω	53	5	85	U	117	u
22	Π	54	6	86	V	118	v
23	Ψ	55	7	87	W	119	w
24	Σ	56	8	88	X	120	x
25	Θ	57	9	89	Y	121	y
26	Ξ	58	:	90	Z	122	z
27	<ESC>	59	;	91	Ä	123	ä
28	Æ	60	<	92	Ö	124	ö
29	æ	61	=	93	Ñ	125	ñ
30	β	62	>	94	Ü	126	ü
31	É	63	?	95	§	127	à

Figure 10 a) and b) ERMES character sets

In this article, the ERMES standard will be referred to as “the ETS” or “the standard”. In presenting the ERMES system, the focus is put on the radio sub-system.

ERMES differs from POCSAG at the service level. While POCSAG offers tone-only, numeric and alphanumeric paging to single users only within home networks, ERMES has several extended features of which the most important are the possibility of roaming between networks and message numbering.

Roaming (i.e. the possibility for a user to have pages redirected to a geographical area different from the home service area for a specified period) is probably the most interesting feature. However, this is not only a technical issue, interoperator agreements are also necessary.

The system capacity of ERMES is 16 channels with a net information rate per channel of approximately 3.75 kbit/s, far more than a single channel in POCSAG systems can provide.

4.1 Migration towards ERMES – MoU

In order to support the start-up of ERMES as a service, and to promote equipment development, the operators of the CEPT countries signed a Memorandum of Understanding (MoU) for ERMES. The MoU is a document where start-up dates and coverage expectancies are lined out and where the signatories oblige themselves to make interoperator agreements about roaming. It is however important to note that the MoU is only an intentional agreement, and does not put any legal constraints on the signatories. At the end of 1994, 33 operators in 21 countries have signed the MoU. This includes even one non-European operator.

The MoU for ERMES basically states the following schedule for implementing the ERMES service:

- To have a commercial start of the service with a minimum coverage of the capital city, or an alternative city, and its corresponding airport(s) by 31 December 1992.

- To have a coverage of at least 25 % of the population or the three largest cities in the country by 31 December 1993.

No operator managed to achieve this, due to a number of factors, which are discussed in some detail in section 4.5, “Status and problems in ERMES”. The first commercial service was launched in France by Infomobile using the service name “Kobby” in November 1994.

4.2 Service aspects of ERMES

ERMES provides different service levels called basic and supplementary services. The most significant are listed and explained below.

4.2.1 Basic services

The following message types are possible in ERMES:

- *Tone-only paging.* The possibility of sending one of 8 different alert types. The alert type shall be distinguishable at the receiver by using e.g. different alert tones.
- *Numeric paging.* The possibility of sending messages containing a maximum of 16,000 characters from the numeric character set defined in the standard and shown in Figure 10.
- *Alphanumeric paging.* The possibility of sending messages containing a maximum of 9000 alphanumeric (text) characters from the character set defined in the standard and shown in Figure 10.
- *Data transmission.* The possibility of sending bit-transparent messages. The maximum message size is 64 kbits.

Tone-only, numeric and alphanumeric paging is mandatory for the network operator, while transparent data is optional. A network operator may set the maximum message lengths shorter than the system limitations listed above, however, a minimum of 20 character numeric and 400 character alphanumeric messages should be supported. Receivers will come in four types, corresponding to the four message types, however, a so-called numeric receiver must also include tone-only reception possibility and alphanumeric receiver must include both tone-only and numeric. Transparent data receivers are special type receivers and do not have to comply with the other message types.

4.2.2 Message numbering and loss of call indication

A new feature in ERMES is the message numbering facility which is a mandatory feature. A message counter is assigned to each subscriber. Messages are numbered cyclically from 1 to 31, and the paging receiver shall be capable of keeping track of incoming numbers. The paging subscriber may then be warned about missing messages and should have the possibility of ordering message retransmission on a per call basis.

4.2.3 Receiving calls outside the home service area – roaming

In ERMES roaming is defined in the following way [5]:

> *The possibility of directing calls to another geographical area than the service area for a time period. Roaming is possible both within the home network and to external networks.*

ERMES defines different “areas” when service and coverage are discussed. The basic entity is the *paging area,* abbreviated *PA*, which constitutes the minimum area to which a paging subscriber is permitted to subscribe in order to receive paging messages.

A *geographical area – GA* – consists of a number of PAs in an operator network. GAs are defined by agreements between network operators for internetwork roaming. It can also be used by a single operator for roaming within his own network.

A *service area* is thus the PAs to which the paging subscriber has subscribed and in which paging messages will normally be transmitted.

Likewise, a *roaming area* can be defined as the GA where the paging subscriber asks for message delivery when using the roaming service.

4.2.4 Group calls

Group calls may be generated in two ways:

- A group of paging receivers sharing a common address.
- Using common temporary addresses (CTA).

When several receivers share the address the call processing is not different from calls to single users from a network point of view.

In the latter case, the procedure is different. First, each pager in the group is individually addressed and messages containing a pointer to a CTA is sent to each of the pagers. The CTA now functions as a valid address and, second, the actual message is transmitted once directed to the CTA instead of the individual addresses.

This technique enables groups to be defined, redefined and deleted in the network only, and pagers do not have to be reprogrammed.

4.2.5 Security aspects

Security contains three aspects: authentication, legitimisation and encryption:

- *Authentication mechanisms:*
  - *by password.* Authentication of the subscriber requesting a password known only by the correct registered subscriber. The registered subscriber can at any time change the password.
  - *by reverse calling.* The authenticity of the subscriber is ensured by disconnecting the call immediately after identification and a new call is set up in the reverse direction.
  - *by use of a certificate.* This is the use of a special authentication certificate, e.g. a smart card in a special terminal.
- *Legitimisation code.* The use of a legitimisation code by the calling party to prove that he is authorised to carry out a particular restricted operation. This code may be known to several calling parties, and is not to be confused with the password which is used to identify a subscriber. An example of such restricted operations may be access to pager members of closed user groups. The subscriber controlling the restricted operation may change the legitimisation code.
- *Encryption of messages.* The possibility of protecting message transmission on the air interface by the use of an encryption algorithm. The transparent data facility is used to transmit such messages.

4.2.6 Special subscriber features

Apart from the previously described services, both paging and fixed subscribers may have access to special features:

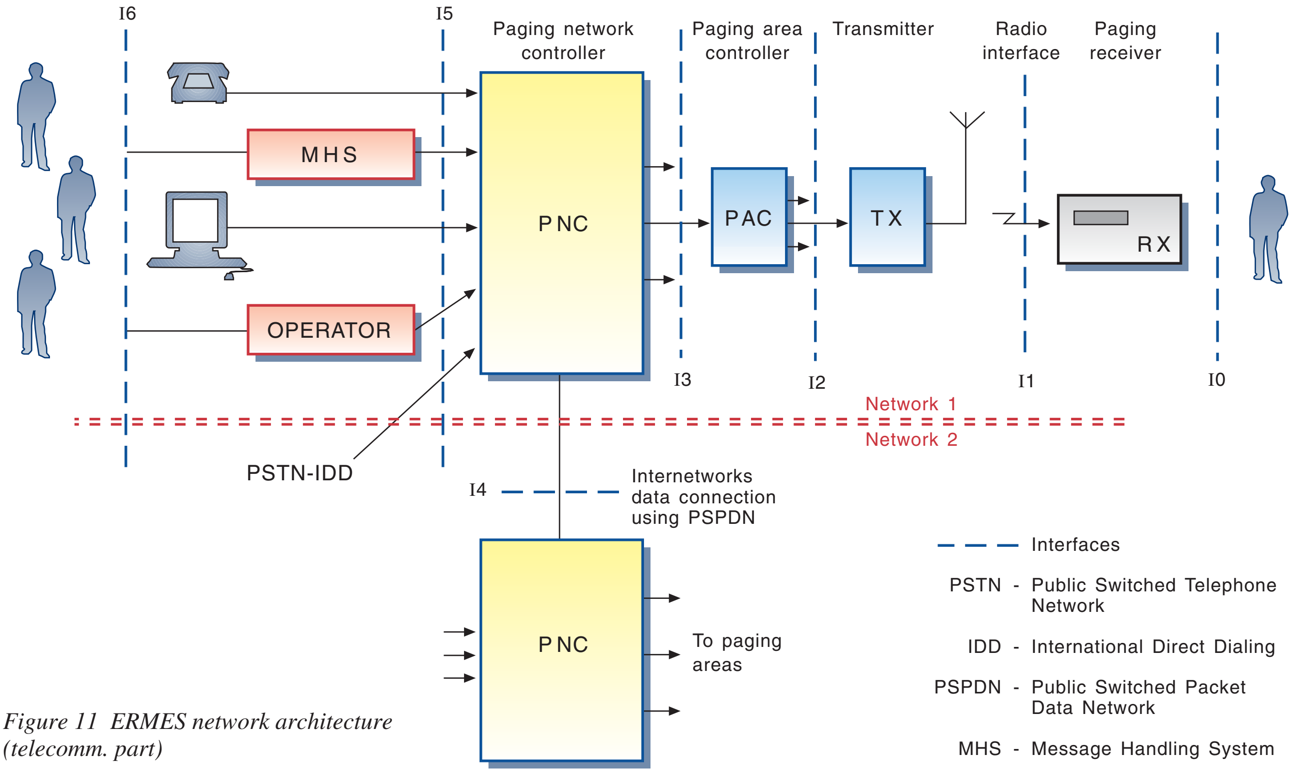

I6
I5
Paging network controller
Paging area controller
Transmitter
Radio interface
Paging receiver
MHS
OPERATOR
PNC
PAC
TX
RX
I3
I2
I1
I0
Network 1
Network 2
PSTN-IDD
I4
Internetworks data connection using PSPDN
PNC
To paging areas
Interfaces
PSTN - Public Switched Telephone Network
IDD - International Direct Dialing
PSPDN - Public Switched Packet Data Network
MHS - Message Handling System

Figure 11 ERMES network architecture (telecomm. part)

- *Temporary barring.* Allows the mobile subscriber to temporarily avoid receiving any calls regardless of their origin.
- *Diversion.* The possibility to have calls diverted to another receiver registered in the same network or in another ERMES network or to another telecommunication system for a period of time.
- *Deferred delivery.* The possibility to instruct the network controller that any message being submitted should be delivered no sooner than at a specified date and time.
- *Directed calls or choice of destination.* The possibility of the calling party to choose the geographical areas that he wants the calls to be transmitted in.
- *Automatic repetition of messages.* Both the calling party and the paging subscriber have the possibility of requesting automatic message repetition. For the paging subscriber this is available on a registration basis, i.e. as part of the subscription, thus affecting all calls. The calling party may order this on a per call basis.

4.3 The ERMES network

An operator network consists of two main components, the telecommunications network part and the operation and maintenance (O&M) network part. Only the functional architecture of the telecommunications network part will be briefly presented here. A comprehensive specification of both parts is of course given in the ETS [6].

4.3.1 System architecture

The functional organisation of the telecommunication network of ERMES is shown in Figure 11. The main components of the network are the *Paging Network Controller* (PNC), the *Paging Area Controllers* (PAC) and the *Base Stations* (BS).

Six system interfaces are identified in the network architecture denoted I1 to I6. I1 is the radio interface between base station and paging receiver which is described in section 4.4, “The ERMES air interface”. The others are part of the network architecture.

The PNC controls the network and manages I3, I4 and I5 interfaces. It performs as the user’s access through the telecommunications access network, it manages and controls the subscriber’s data base and it controls the radio transmissions in the areas under its responsibility.

The PNC interacts with other networks through the I4 interface. The PNC can control up to 64 PACs through the I3 interface.

The main task of the PAC is to organise paging message queuing and batching. It also manages the priority of paging messages. Paging messages are delivered to the BSs through the I2 interface.

The BS manages the I1 interface through the transmitters.

4.3.2 Access methods

Several access methods are available as input to ERMES paging, both telephonic and non-telephonic. Telephonic access obviously makes use of an ordinary telephone set (DTMF), while non-telephonic means access through any other type of terminal.

Both telephonic and non-telephonic access comprises two methods; the one stage and the two stage selection method. In the one stage method, the Address Code (AdC) of the pager subscriber is entered directly through the terminal, while in the two stage method, the service number (SN) is entered first and then the AdC when connection with the PNC is established. The understanding of AdC and SN will become clearer when telephonic access is described. Only telephonic access will be described in some detail.

4.3.2.1 Telephonic access

Telephonic access requires a DTMF telephone set. Tone-only, numeric and alphanumeric input is possible. Both one stage and two stage methods may be used. In the one stage method, the AdC is simply the pager's telephone number. In the two stage method, the SN is the telephone number to the service provider, and the AdC is entered on the telephone keyboard after some greeting message in which instructions are often given.

Tone-only access is straight forward, as only the AdC or SN + AdC is entered to generate a call.

Numeric calls require the caller to also enter the message. The non-numeric characters shown in Figure 10 are mapped on the telephone keyboard as follows:

'/'	= *1;	<space>	= *2;
'U' (urgent)	= *3;	'-' (hyphen)	= *4;
'.' (full stop)	= *5;	'%'	= *6.

Thus, the special characters are entered by preceding the proper number key with the asterisk key (*) as a shift-key. For example the numeric message 47/63-80-91-00 is entered: **47 *1 63 *4 80 *4 91 *4 00**.

Entering alphanumeric messages becomes more complicated as more characters are required. Not all characters from the set shown in Figure 10 are available from a DTMF telephone, only the numeric characters and the capital letters A – Z. The keyboard mapping is defined by the operator, and one proposal is the CEPT T/CS shown in Figure 4.

4.3.2.2 Non-telephonic access

Non-telephonic access is possible through a number of different access networks. The following are defined in the standard [6]:

- Alphanumeric terminal access, e.g. by the use of a modem and UCP.
- Telex access, the paging subscriber is assigned a telex number within the national telex numbering plan.
- Message Handling System (MHS) access, e.g. X.400 based e-mail systems.
- Bureau access, i.e. access via an operator.
- Videotex access, i.e. through an F.300 based Videotex service. A videotex terminal is assumed.
- ISDN access.

4.4 The ERMES air interface

The ETS Part 4, Air interface specification [7] contains the description of the radio resource utilisation, and is the part of the standard most of all affecting the quality of service. Several contradictory objectives were to be fulfilled at the same time:

- Good spectrum efficiency
- Secure and robust transmission
- Small end-to-end delay
- Cheap user equipment
- Good pager battery economy.

The last objective is not the least, and a great deal of effort was put into achieving this. It does not only have a general battery saving effect, but this also opens the possibility of integrating pager receivers in small equipment.

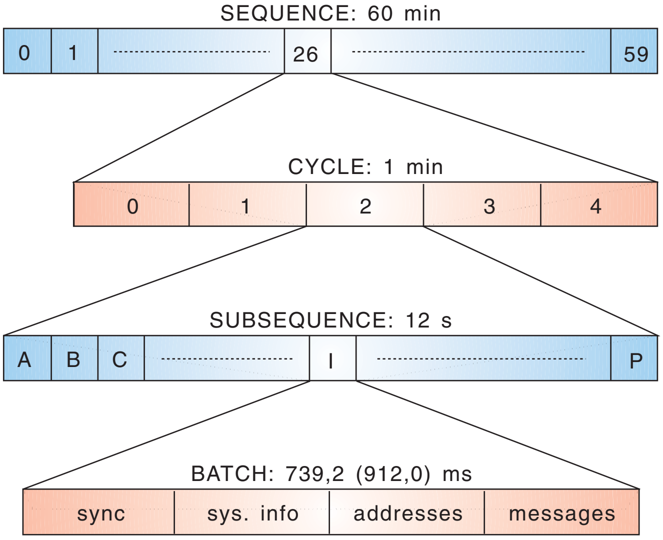


Figure 12 The ERMES radio protocol on RF channel 0

4.4.1 Overview

The ERMES air interface specifies the format of the transmitted signal from the base station transmitters, i.e. the radio protocol, error correction techniques and RF-modulation format.

ERMES is allocated the VHF-frequency band from 169.4125 MHz to 169.8125 MHz. The band is divided into 16 channels, each occupying 25 kHz. The channels are numbered from 0 to 15, and the channel centre frequencies are given by the equation

$$f_n = 169.425 + n \cdot 0.025 \quad [\text{MHz}] \quad (4.1)$$

Thus, channel 0 has a centre frequency of 169.425 MHz and channel 15 has 169.800 MHz. ERMES pagers have to be able to receive calls on all 16 channels. In order for the pager not to have to scan all channels at all times (and consuming battery power at a non-acceptable level) network operators are advised to use some or all of the battery saving techniques built into the protocol. This is also an advice to pager manufacturers to make use of the information in the transmitted signal to build "intelligent" pagers.

The radio transmission uses a four level modulation method called 4-PAM/FM (4-level pulse amplitude modulation/frequency modulation), with a signalling rate of 3125 baud. This gives a gross information rate of 6250 bit/s.

A forward error correction scheme using a (30,18) modified BCH [15] block code is used to add redundancy to the transmitted signal. This is combined with an interleaving technique performed on blocks of nine code words.

The radio protocol is hierarchical in four levels, *sequence, cycle, subsequence* and *batch.* Furthermore, the batch is divided into four distinct partitions: *synchronisation partition, system information partition, address partition* and *message partition*. An overview of the radio protocol is given in Figure 12.

To secure that a pager will be able to receive calls on all channels, a channel synchronisation is needed. Each pager will be programmed to belong to a specified batch, A to P. Thus, the 16 channels are inter-synchronised such that the first batch of the subsequence is different on all the channels. 16 batches distributed on 16 channels, secures that a batch of type e.g. "B" is never sent simultaneously on two or more channels. Batches

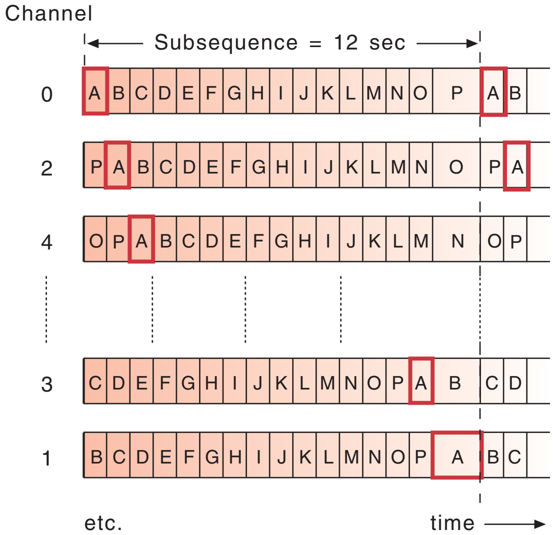

Channel
Subsequence = 12 sec
0
A B C D E F G H I J K L M N O P A B
2
P A B C D E F G H I J K L M N O P A
4
O P A B C D E F G H I J K L M N O P
3
C D E F G H I J K L M N O P A B C D
1
B C D E F G H I J K L M N O P A B C
etc.
time

Figure 13 ERMES channel synchronisation

of the same type are sent on the channels in the following order: 0-2-4-6-8-10-12-14-15-13-11-9-7-5-3-1. Figure 13 illustrates the inter-channel synchronisation scheme.

Note that only batches are rearranged and that subsequences always start simultaneously for all channels. Thus, for channel 0, the subsequence starts with batch A, while for e.g. channel 8, the subsequence starts with batch M.

In the standard the air interface transmissions are organised in four levels:

- *L4 – information format.* This is the basic coordination of the paging system data and paging message data. i.e. how the transmitted data is organised in a way that can be recognised by the pager receiver.
- *L3 – Error correction coding.* Forward error correction (FEC) is used to add redundancy to the transmitted data.
- *L2 – Code word interleaving.* In addition to FEC, interleaving is used to spread burst errors, which may occur in a mobile environment. This increases the effect of the error correction.
- *L1 – Modulation.* This level describes how the coded and formatted data are distributed from the base station transmitters by radio transmission.

4.4.2 Pager addressing structure

Every pager has a unique address, denoted radio identity code – RIC. The full RIC consists of 35 bits, containing the following information:

- Operator identity – OPID – 13 bits, consisting of:
  - Zone code – 3 bits
  - Country code – 7 bits
  - Operator code – 3 bits
- Local address – 22 bits, consisting of:
  - Initial address – IA – 18 bits
  - Home batch number – 4 bits.

Pagers are addressed and messages sent using a method called *repeated addressing*. First, the pager initial address is sent. This “wakes up” the pager. The actual message is sent at a later time, not more than 12 seconds delayed, starting with a repetition of the IA before the actual message contents.

The zone and country code follows CCITT recommendation E.212 (1988) which is included in the ETS Part 4.

4.4.3 Information format

The basic entity of the transmission is the batch. It is organised in four partitions as shown in Figure 14. The partitions all have different tasks, and carries different information.

4.4.3.1 The synchronisation partition

Two 30 bit words, the preamble (PRE) and the synchronisation word (SW) may be used by the pager to obtain bit- and code word synchronisation, respectively. They are fixed pattern words defined as:

PRE:
00 10 00 10 00 10 00 10 00 10 00 10 00 10 00

SW:
10 00 10 10 00 10 00 00 10 10 00 00 10 10 10

The PRE word is chosen to be alternating symbols, 00 and 10, which corresponds to the two outermost frequencies in the 4-PAM/FM modulation scheme used. (This is described in more detail in section 4.4.6, “Modulation”.)

The SW word also consists of the two outermost symbols and is chosen to have good autocorrelation properties and large Hamming distance to all other code words used in the protocol.

4.4.3.2 The system information partition

Three code words contain data about network operator; cycle, subsequence and batch number and information relevant to visiting receivers from other network.

4.4.3.3 The address partition

This partition is of variable length with a maximum of 140 code words. 18 bit initial addresses (IAs) are transmitted in descending order and the address partition is always terminated with a special

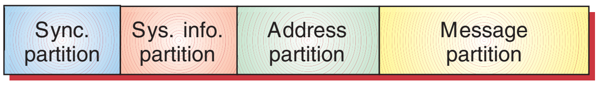
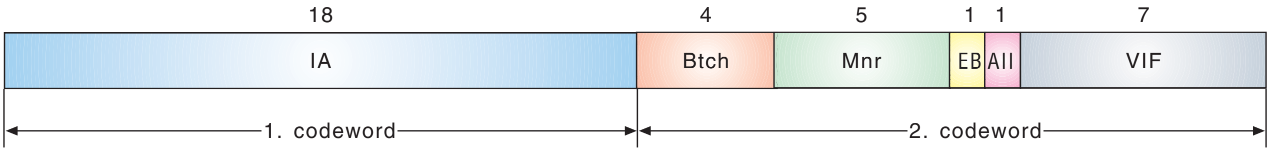


Figure 14 Batch structure



- IA (Initial address, 18 bit) + Btch (Home batch number, 4 bits) = Local address, 22 bits
- Mnr = Message number, 5 bits
- EB = Extemal Bit, 1 bit
- AII = Additional Information Indicator, 1 bit
- VIF = Variable Information Field, 7 bits

Figure 15 ERMES message header (MH) format

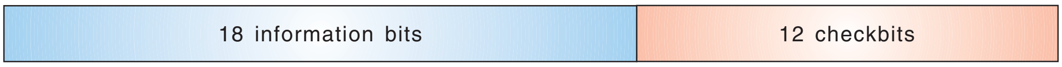


Figure 16 (30,18)-code word structure

30 bit "illegal" word, the address partition terminator (APT) defined as:

10 01 00 11 10 00 01 10 00 10 00 11 10 01 00

4.4.3.4 The message partition

The normally largest and remaining part of the batch is used to transmit the actual message contents. The code words are assembled in code blocks of 9 code words each. Each code block is sent to the modulator interleaved (see section 4.4.5, "Code word interleaving").

Each message starts with a message header of 36 bits, fitting into two code words, the contents of which is shown in Figure 15.

As the local address is made up of the IA and home batch number, the first code word is the initial address repeated. The message header may in fact consist of more than two code words if the Additional Information Indicator (AII) is set to one, however, these features will not be treated in detail here. The contents of the variable information field (VIF) is also dependent on the AII, and contains, among others, information about *message type* (tone, numeric, alphanumeric or data) and *alert type* (8 different).

Immediately after the message header, the message body is transmitted. The message contents are coded differently according to the message type. Tone-only messages (type 0) has no body, thus no code words are needed for this. Numeric messages (type 1) are coded using a 4-bit alphabet, while alphanumeric messages (type 2) uses a 7-bit alphabet designed specially for ERMES. Both these character sets are shown in Figure 10.

The characters are packed one after another into the code words. The remaining part of the last code word is filled with terminating characters, which for numeric messages are the "space"-character (1011), while for alphanumeric messages it is the "DC1"-character (0010001), also referred to as the *End Of Message* (EOM) character.

The fourth message type (type 3) is the transparent data, thus the alphabet is the binary code zeros and ones. Using this, free format data may be transmitted.

Each message is preceded and succeeded by a special word, the *Message Delimiter* (MD), however such that not more than one MD is transmitted between messages. The MD is the fourth of the "illegal" code words defined as:

11 01 01 01 11 10 01 11 11 10 11 10 11 10 11

When a batch is not completely filled with addresses and messages, the remaining part may be filled with MDs to complete the transmission or the transmitters may be turned off until the start of the next batch. In any case, the last code block containing message information must be completed with MDs and transmitted.

4.4.4 Error correction coding

The term code word has already been introduced when describing the information format. The transmission is protected using a (30,18) block code. This code is based on the well known (31,21) BCH-code [14], the same code which is the basis in POCSAG.

The (30,18) code is a systematic block code of 18 information bits and 12 check bits. The code word structure is shown in Figure 16.

The code words may be generated using the generator polynomial *g(X)* on the information polynomial *m(X),*

$$c(X) = m(X)X^{12} + m(X)X^{12} \bmod g(X) \quad (4.2)$$

where

$$g(X) = X^{12} + X^{11} + X^9 + X^7 + X^6 + X^3 + X^2 + 1 \quad (4.3)$$

$$m(X) = m_{17}X^{17} + m_{16}X^{16} + m_{15}X^{15} + \ldots + m_1X + m_0 \quad (4.4)$$

The coefficients m_0 to m_{17} represent the 18 information bits to be coded.

This code has a Hamming distance of 6, which makes it capable of correcting any two errors per code word using hard decision, and detecting any three errors. Additionally, several other error patterns containing more than two errors can be corrected.

4.4.5 Code word interleaving

Paging is mobile communications and as such subject to a constantly changing radio channel, usually referred to as multipath fading. Fading causes transmission errors to occur in bursts, often longer than the block code can handle. The overall bit error rate may however be small enough for the code to correct if errors were evenly distributed. In order to utilise the code better, and thus protecting the transmission, bit-interleaving is used on the message partition of the batch. In this way errors are spread over several code words, making them correctable.

In ERMES, an interleaving depth of 9 is used. This means that 9 code words are collected to form a matrix of 30 x 9 = 270 bits, called a *code block*. This operation is performed on the transmitter side immediately before the bits are sent to the modulator. Bits are then read out from this matrix column by column starting at the MSB of the first code word as Figure 17 shows.

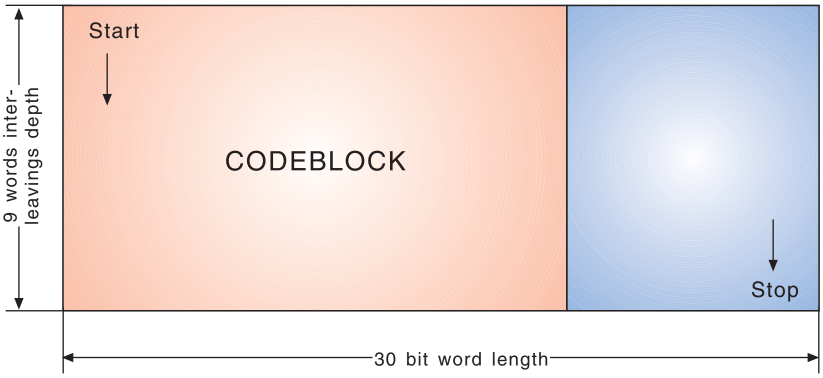


Figure 17 Code word interleaving in ERMES

4.4.6 Modulation

The modulation technique used in ERMES is called 4-PAM/FM standing for four-level pulse-amplitude modulation/frequency-modulation. In fact, this is simply a 4-frequency shift keying method with continuous phase. The frequency shifts are smoothed using a baseband filter.

As a 4-level modulation method, dibits are coded into symbols using the following symbol alphabet:

Nominal frequency:	*Symbol*:
carrier + 4687.5 Hz	10
carrier + 1562.5 Hz	11
carrier – 1562.5 Hz	01
carrier – 4687.5 Hz	00

The data rate fed into the modulator is 6250 bit/s, giving a *symbol rate* "on the air" of 3125 baud.

The frequency shift smoothing is performed by a premodulation filter. This filter is given as an upper and a lower mask. Thus, different practical filters may be implemented as long as their transfer functions are between the mask values. One such filter is a 10th order Bessel filter with 3 dB bandwidth of 3.9 kHz.

This filter is indirectly specified in Part 6, Base station specification [9], through the constraint of the symbol transition shaping. The rise time (10 – 90 %) is set to be 88 µs ± 2 µs, which is a consequence of the frequency mask.

4.4.7 Operation of the radio subsystem

ERMES offers great flexibility in how the radio transmissions are organised to utilise the radio spectrum for maximum coverage. As in POCSAG networks, ERMES planning is based on simulcast coverage. Additionally, both frequency and time division between paging areas (PA) may be used.

Frequency divided networks use different frequencies (ERMES channels) for adjacent paging areas, thus providing the total capacity of the channel for one paging area.

Time divided networks use the same frequency in adjacent PAs. The time division is performed on the subsequence level, i.e. up to 5 adjacent PAs may share the same channel. This offers a great flexibility in allocating resources between PAs as there are several ways of splitting five subsequences of a cycle between e.g. three paging areas as illustrated in Figure 18.

Both frequency and time division may be combined in the same network with some limitations.

4.4.8 Battery saving techniques

One of the great advantages of ERMES compared to previous paging standards is the extended battery saving capability built into the radio protocol. This topic is extensively covered in the ETR [13], and just a brief overview is given here.

Most of the time a pager spends its life not receiving messages, thus it may be sufficient to listen to only 7 or 8 code words per batch to determine that no pages will be transmitted to it during this batch.

Since the radio protocol has four hierarchical levels, it is possible to assign pagers to listen to parts of the transmission only. Three levels of battery saving are defined:

- Subsequence assignment (mandatory)
- Cycle assignment (optional)
- Sequence assignment (optional).

The lowest level is the subsequence assignment of pagers. As described when outlining the pager addressing structure, each pager is assigned a "home" batch through the four LSBs of the pager's local address. A particular pager thus only has to listen for addresses in batches of its own type giving a battery saving ratio (BSR) of 1 : 16.

Additionally, it is possible to program a pager to listen to only a subset of the transmitted subsequences of a cycle. Maximum BSR due to cycle assignment is 1 : 5.

Lastly, it is possible to program the pager to listen to only a subset of cycles within a sequence. A maximum BSR due to sequence assignment is 1 : 32.

Using all these features it is possible to have a total BSR of 1 : (16 x 5 x 32) = 1 : 2560, however, it is not likely that a customer would like to wait more than half an hour to receive a page, thus sequence level assignment will probably at a maximum be used to a level of every 2nd cycle. Still, an excellent BSR is achievable.

As the two latter techniques are optional, pagers will initially be programmed to listen to all subsequences and all cycles. Cycle and sequence assignment is only meaningful as a part of the service provided by the operator, as traffic has to be packed in the assigned subsequences and cycles for the pager to receive them at all.

4.5 Status and problems in ERMES

For several reasons the operators have been reluctant to implement an ERMES service, but as mentioned earlier the first commercial service was opened in France in November 1994. There are now two operators in France. TDR opened its service "TAM TAM" at the beginning of 1995.

Telenor Research, in cooperation with Telenor Mobil, developed an ERMES test system with four base stations and testing was started in March 1991. These were the first actual full scale ERMES transmissions "on air". Later, the network was used to perform tests and measurements to evaluate different aspects of the standard. A possibility to route pages from Telenor Mobil's alphanumeric POCSAG service, "PS-tekst" into the ERMES test system was implemented, and though an ERMES service has not been offered commercially, we can call this the first operating ERMES network in Europe.

The far most important reason for the slow start, is probably the availability of user equipment. So far, the telecommunications industry has not been very eager to develop and start marketing ERMES receivers. One vendor has developed a pager sponsored by the MoU-group and another vendor has announced an

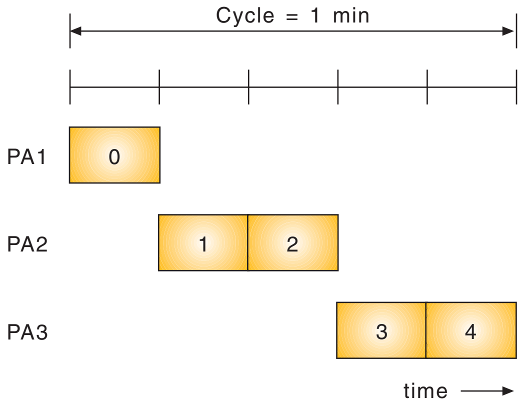


Figure 18 Time division on the subsequence level

ERMES pager. The MoU-group, consisting of operator companies, thought this could give a boost to a rapid development of pager equipment. The fact that several operators obliged themselves to more or less guaranteeing a minimum coverage at a given time, was thought to have such an effect. This has not happened, partially because the operators themselves have regarded the MoU to be less obliging probably due to some of the factors mentioned below, and maybe others as well.

Another reason is the fact that the existing national paging networks provide a good service as almost all MoU countries have had POCSAG networks for some years. They have good coverage and provide the necessary national services, both numeric and text. The roaming possibility has not been considered important enough to make the necessary investments. As POCSAG networks are not standardised otherwise than the radio code, operators have installed different supplementary services such as group messaging and broadcast services.

While ERMES is suffering from a slow start, the new pan-European mobile telephone system GSM is being implemented very rapidly throughout Europe. The GSM short message service will basically offer the same service as a conventional paging network in addition to the fact that acknowledgement from the end recipient is possible. It is very difficult to foresee whether GSM SMS will be a real competitor to conventional paging. It will surely depend on several factors, as for example user equipment size, cost and battery economy, coverage and subscription fees.

It is not only economic and commercial aspects that prevent the introduction of ERMES. An interference problem with television broadcasting has been encountered. Several European countries use the television band III i.e. VHF channels 5 – 12 for ground based television broadcasting and cable television installations. An interference problem has been detected on TV-channel 5 (TV E-5) for both broadcast and cable television. TV-channel 5 has the following nominal carriers:

- Vision carrier at 175.25 MHz (176.00 in France)
- Sound carrier at 180.75 MHz (182.50 in France)
- Second sound carrier (for analogue stereo) at 180.992 MHz
- NICAM carrier at 181.10 MHz.

Due to problems encountered in Germany, CEPT set down a joint project team to study the compatibility problems between ERMES and TV E-5 [16]. Their report concludes that two major mechanisms are the sources of this disturbance:

- ERMES channels occur as image frequencies of the TV sound carrier with respect to the vision carrier in the IF stage of the TV receiver.
- One particular intermodulation product falls in the range 180.7 – 181.075 MHz.

The first effect is the dominant interference mechanisms in TV receivers with poor IF selectivity, a problem which can be corrected by the use of in-line filters in front of the active components of the TV receiver. The interference problems have so far lead the German regulatory authorities to reject to give out ERMES licenses to any operator.

5 References

1 Radiopaging Code Standards Group. *The book of the CCIR radiopaging code no. 1.* Secretary RCSG, British Telecom, 1986.

2 CCIR Rec. 584-1. *Standard codes and formats for international radio paging.* In: Recommendations of the CCIR, 1990, vol. VIII - Section 8A, 11–15.

3 ISO. *Information processing : ISO 7-bit coded character set for information interchange.* Geneva, 1983. (ISO 646-1983.)

4 European Telecommunications Standards Institute (ETSI). ETS 300 133-1: *Paging Systems (PS); European Radio Message System (ERMES) Part 1: General aspects.* Sophia Antipolis, 1992.

5 European Telecommunications Standards Institute (ETSI). ETS 300 133-2: *Paging Systems (PS); European Radio Message System (ERMES) Part 2: Service aspects.* Sophia Antipolis, 1992.

6 European Telecommunications Standards Institute (ETSI). ETS 300 133-3: *Paging Systems (PS); European Radio Message System (ERMES) Part 3: Network aspects.* Sophia Antipolis, 1992.

7 European Telecommunications Standards Institute (ETSI). ETS 300 133-4: *Paging Systems (PS); European Radio Message System (ERMES) Part 4: Air interface specification.* Sophia Antipolis, 1992.

8 European Telecommunications Standards Institute (ETSI). ETS 300 133-5: *Paging Systems (PS); European Radio Message System (ERMES) Part 5: Receiver conformance specification.* Sophia Antipolis, 1992.

9 European Telecommunications Standards Institute (ETSI). ETS 300 133-6: *Paging Systems (PS); European Radio Message System (ERMES) Part 6: Base station specification.* Sophia Antipolis, 1992.

10 European Telecommunications Standards Institute (ETSI). ETS 300 133-7: *Paging Systems (PS); European Radio Message System (ERMES) Part 7: Operation and maintenance aspects.* Sophia Antipolis, 1992.

11 ITU-R. Recommendation ITU-R M.539-3. *Technical and operational characteristics of international radio-paging systems.* Geneva, 1994.

12 CCITT Q.400 - Q.490. Specification of signalling system R2 : recommendations Q.400 - Q.490. In: *CCITT Blue Book vol VI - fascicle VI.4,* 33-143.

13 European Telecommunications Standards Institute (ETSI). TBR 7: *Paging Systems (PS): European Radio Message System (ERMES) Receiver requirements.* Sophia Antipolis, 1993.

14 European Telecommunications Standards Institute (ETSI). ETR 050: *Paging Systems (PS); European Radio Message System (ERMES).* Sophia Antipolis, 1993.

15 Lin S, Costello D J. *Error control coding : fundamentals and applications.* N.J., Prentice-Hall, 1983. ISBN 0-13-283796-X.

16 European Radiocommunications Committee (ERC). *Radiocommunications reports. ERMES / TV E-5 Compatibility.* Montreux, October 1993. (ERC Report 22.)

MOBSIM – a computer simulation program for the analysis of the traffic handling capabilities in a GSM network

BY FINN TROSBY, STEIN SVAET, TORE J BERG, HOGNE SOLVOLL AND JON MARTIN FURSETH

1 Introduction

As a result of the world-wide liberalisation of telecommunications services provision, all network operators have focused on the problem of how to get the most out of their investment, and how to plan, implement and run their networks in an optimal way. This also applies within the area of mobile services provision, where the liberalisation first begun and the competition is at its peak. GSM, the first mobile communication system genuinely developed for competition, offers the operator a huge set of parameters for optimal tuning of their networks. Each operator knows that the question of whether the services provided have a high technical quality, is crucial, where the term 'service quality' is taken in a very wide perspective. For providers of mobile services, the question of technical standards is perhaps even more important than for fixed network operators. This is so because acceptable radio coverage provision in vast areas and for many types of environments is hard to accomplish, and because the mobile users tend to expect and demand the same quality of the mobile services as of those offered within the fixed networks.

Efficient tools for detailed planning of mobile networks are therefore definite necessities for the ambitious provider of mobile services. Within this toolbox MOBSIM finds its place as a measure to ensure the traffic handling capability of a specific GSM network alternative for a certain area. MOBSIM is a computer simulation program for the analysis of the traffic handling capabilities in a GSM network, and was developed during a two-years period upon overall requests and guidance from Telenor Mobil. MOBSIM is to be run on UNIX based workstations. The programming language employed was C++.

Normally, the mobile network planning department starts off with purchasing digital maps of the areas to be covered. These data are then fed into a CPS program, which produces estimates of path loss of all area segments according to a pre-defined geographical resolution. In certain areas, e.g. along important highways with very dense car traffic, the path loss estimates are replaced by more reliable measurements. These data, being either CPS estimates or measurements, are fed into MOBSIM together with data describing the mobile users with respect to number, moving habits, call rates and call repetition behaviour. Also the planner's alternative of how to allocate the 200 kHz spaced GSM carriers between the different base stations, his choice of parameter values of the handover and power control algorithms along with his decisions on e.g. umbrella cell features are fed into MOBSIM before starting the simulation. Output from the MOBSIM simulation is traffic characteristics such as carried traffic and congestion values, quality of service figures like number of dropped calls, and more system specific characters like C/I statistics and handover rates. Thus, MOBSIM is in itself no optimisation tool; the user has to specify a certain case or rather a set of cases. The merits of MOBSIM are its capability – provided reliable input data – to evaluate the cases presented and help the planner to find the most suitable of these.

The above description suits the most intuitive use of MOBSIM. However, MOBSIM provides the benchmark for many functionality tests. For example, when being interested to judge which is the best one of two or more handover algorithms, program code describing the new algorithms is all that is required. The heavy stuff, i.e. the radio channel and protocol modelling, the interference calculations, the description of the mobiles and their mobility patterns, the path loss data structure, the statistical calculations, is there already.

2 Terminology

The terminology in this document conforms to the terminology in the GSM specifications and also to the terminology of [9]. Exceptions from this rule are the mobile stations and the base stations.

In this paper, one distinguishes between the *mobile terminal* (MT) and the *mobile station* (MS). The mobile terminal deals with the functionality related to the telecommunication tasks of the mobile entity. The mobile station deals with the functionality related to the mobility of the mobile entity, e.g. stepping the mobile entity forward along its geographical path.

Also, the entry point at the fixed side together with the radio transceiver and associated equipment is called a *base station* (BS). A base station in MOBSIM is equivalent to a cell in the GSM network with associated functionality in a BTS and a BSC. It might be regarded an entity with a set of radio resources (frequencies and logical channels) and functionality for administration of calls (handover, power control, call control, etc.).

A comment should be made on a certain term *radio data*, which is frequently used throughout this document. Radio data is always related to a geographical location, and denotes the path loss values between the geographical location and the closest base stations. The geographical location may either be represented by a small square or a piece or 'pixel' of the slope of a road.

3 Overall description

The run of a simulation using MOBSIM may be considered a two-step process, see Figure 1.

First there is a so-called trace generation, performed by a separate module, *Trace Generation* (TG), in which traces of all moving mobiles are specified. These traces also comprise information about the offered teletraffic, i.e. indications of at which geographical locations the respective mobile intends to initiate a call.

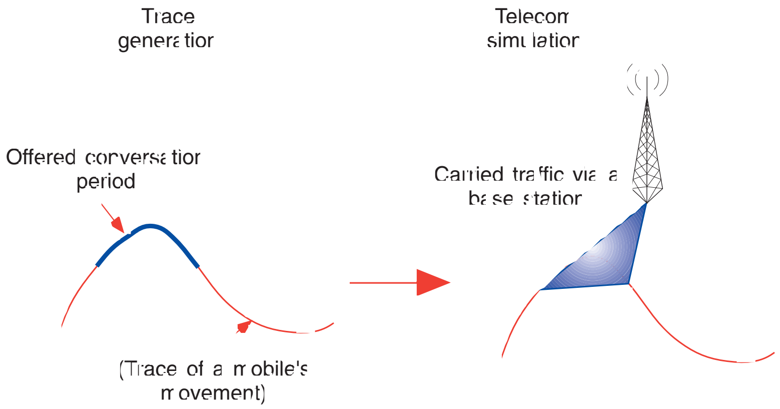

Trace generation
Telecom simulation
Offered conversation period
Carried traffic via a base station
(Trace of a mobile's movement)

Figure 1 The two basic simulation phases comprised by the simulator

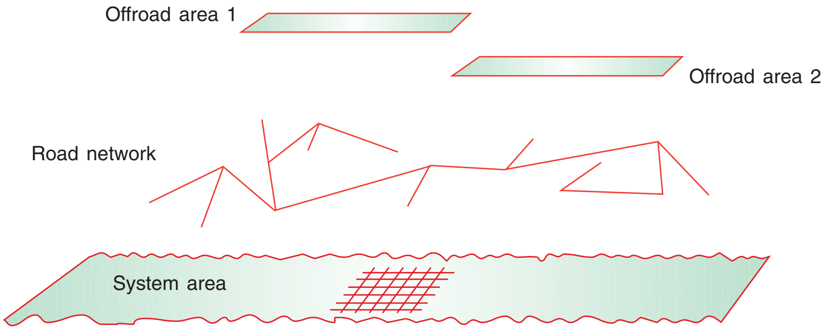


Figure 2 Example of how two offroad areas and one road network may be specified as input for one simulation. The offroad areas and the road network are projected down to the system area

Secondly, the traces produced by TG are processed by another module, *Telecom Simulation*, (TS), in which the carrying of the offered teletraffic by the specified GSM network is simulated. From simulation, the desired information – congestion statistics, C/I statistics, dropped call statistics, handover statistics – will be derived.

As stated above, the task of the TG module is to generate descriptions of the movement of the mobiles and the teletraffic which they offer. This is based on user defined average number of mobile terminals in the system. The call generation is controlled by the distributions of the interarrival and holding times for each service.

A mobile can either move along a predefined track (road) or randomly within an off-road area, see Figure 2.

The offroad areas are characterised by their geographical location, the number of mobiles moving inside each area, and the mobility pattern of these mobiles. The shape of the offroad area is – so far – always rectangular. The mobility pattern of the mobiles is always some sort of brownian motion.

The road network is a graph, and can be either a tree or a mesh network. The mobiles move according to flow figures defined for the road arcs. Like an offroad area, a road network comprises a certain number of mobiles.

There may be several offroad areas and several road networks defined within input data set for one simulation. Note, however, the restriction that a mobile may under no circumstances leave one offroad area or road network to enter another offroad area or road network.

The task of the TS module is to establish the data structures of the mobile telecommunications network and to see how it carries the offered teletraffic. Currently, only the GSM 900 system is modelled for the TS. The main functionality of the TS module is to simulate:

- ***Call Set-up***, for which cell selection and call establishment are evaluated. In the latter case the success or failure of the signalling and the SDCCH-TCH allocation are evaluated. The functionality to make assignments to other cells is implemented.
- ***Handover.*** During the call and call establishment it is possible to make intra- or intercell handover.
- ***Radio Link Control***. The quality of the uplink and downlink is evaluated; the call will be dropped if the quality criteria are violated.
- ***Power Control***. Mobile station power control may be performed.
- ***Measurement Reports***. MOBSIM evaluates measurement reports including RXLEV and RXQUAL for own cell uplink and downlink, RXLEV for other base stations are generated every 480 ms. It is possible to control whether or not only neighbouring base stations shall be included in the reports.
- ***User Behaviour.*** A certain user behaviour may be defined, to characterise e.g. the number of retries when the call is either not established (busy) or dropped, the time between the retries, and duration of spurts and pauses.
- ***DTX.*** Discontinuous transmission, DTX, may be performed for both uplink and downlink and can be switched on/off individually per BS.
- ***GSM Channel Types.*** MOBSIM allocates for each TDMA time slot on each carrier a GSM channel type. MOBSIM assumes that any pair of BSs (in GSM context: BTSs) are fully synchronised. However, parties involved in different connections currently on different GSM channel types but at the same time slot and on the same carrier frequency are assumed to be interfering with one another according to simple average schemes.

4 Handling of geographical data

The geography module in MOBSIM establishes a data structure that describes geographical objects and information that can be associated with or referenced through these. Geographical objects are described using a two-dimensional Cartesian coordinate system.

The geography module serves different purposes in trace generation and in telecom simulation. In TG, it assists in building traces of the moving mobile terminals. In TS, it assists in the retrieval of the signal strength data of the terminal. Information about the radio transmission between the different base stations and different geographical locations (so-called radio data) are then fed to the terminal.

The geography module deals with two entities – *road networks* and *offroad areas*, which have no direct relations with each other. No means is provided for a mobile in one network or offroad area to go to another network or offroad area. The major geography objects modelled are offroad areas and road networks.

4.1 Offroad areas

The offroad area model is used for simulating mobile terminals moving within a closed area. The offroad area is a closed polygon, currently restricted to form a rectangular shape. Common for offroad mobiles is that their movement is not limited to a specific path, but can move around unrestricted within the boundaries of the offroad area.

The notion of offroad mobiles may emulate that of motorised vehicles like cars or buses, or walking mobile users within a park or boats on a lake or in a harbour. In dense road networks, the offroad model could also be used for representing the road network.

4.2 Road networks

The road network model is used to simulate mobile users moving in a network of connected paths. A graph model is used for road networks, so the paths (roads) are arcs, connected by nodes (crossings). The road consists of linear sections, 'road segments'.

As the name indicates, the road network model most frequently applies to real life road networks, but the graph model could also be applied to e.g. footpaths, boat routes or railway lines.

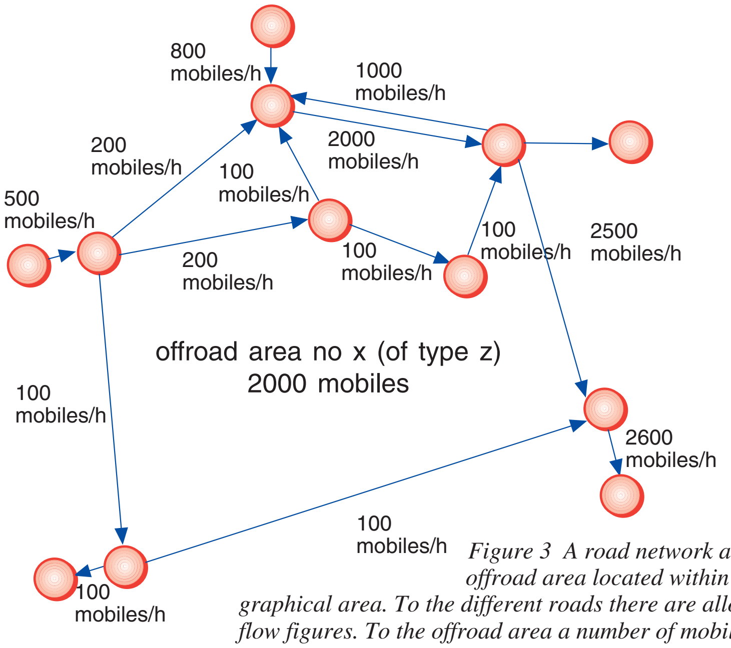
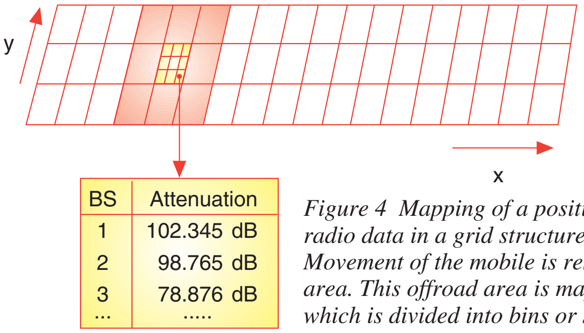


Figure 3 A road network and an offroad area located within a geographical area. To the different roads there are allocated flow figures. To the offroad area a number of mobiles is allocated

4.3 Data

Besides the geographical description of road networks and offroad areas, the geography module also handles data that for obvious reasons is organised together with these, such as speed limits, traffic flow, etc.

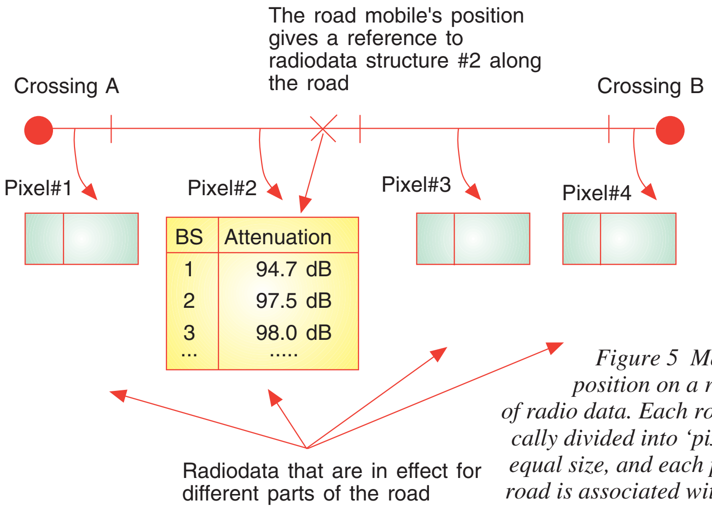
The geography module also supplies the TS with radio data for the locations that the mobile terminal is passing. The radio data can be based on imported coverage predictions or measurements. For the offroad areas, radio data is stored in a grid-like structure that allows multiple resolutions (see Figure 4).



BS	Attenuation
1	102.345 dB
2	98.765 dB
3	78.876 dB
...	.....

Figure 4 Mapping of a position in an offroad area to radio data in a grid structure: Movement of the mobile is restricted by an offroad area. This offroad area is mapped onto a main grid which is divided into bins or sub grids with higher resolution. Attached to each bin is a set of radio data

Each road in the road network is logically divided into "pixels" of equal size, and each pixel is associated with a set of radio data (see Figure 5).


The road mobile's position gives a reference to radiodata structure #2 along the road
Crossing A
Crossing B
Pixel#1
Pixel#2
Pixel#3
Pixel#4
Radiodata that are in effect for different parts of the road

BS	Attenuation
1	94.7 dB
2	97.5 dB
3	98.0 dB
...	.....

Figure 5 Mapping of a position on a road to a set of radio data. Each road is logically divided into 'pixels' of equal size, and each part of the road is associated with a set of radio data. Start and end pixels may be of different size

The major advantage of the road network model over the offroad area model is that the mobiles travel the same path, and access the same set of data. Such data can therefore be cached and available with significantly less processing than for the offroad area. All mobiles follow different routes in the offroad area, which causes a need for individual handling of each mobile.

5 Mobile trace generation

A separate module – TG – is used for generating offered teletraffic and the mobility patterns of the mobiles in the simulation area. The output from this module is 'traces', being used as input to the TS. The trace describes the call and the mobile's movement during the call period, including an additional trace extension that allows re-establishing an interrupted call and retry if the call set-up procedure fails.

A basic assumption of the model is that *the teletraffic conditions do not have any impact on the mobility behaviour* of the mobile stations. (However, mobility conditions may have considerable impact on the offered teletraffic.) It is also assumed that the desired teletraffic load can be described as a linear function of the car or pedestrian traffic flow.

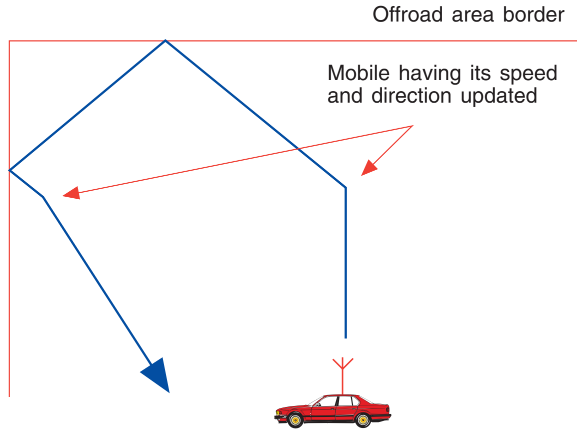


Figure 6 Offroad mobile's movement pattern. MOBSIM allows for defining how often the mobile's linear movement may be updated or changed. When encountering the offroad area border, the mobile is reflected. MOBSIM allows for several reflection schemes

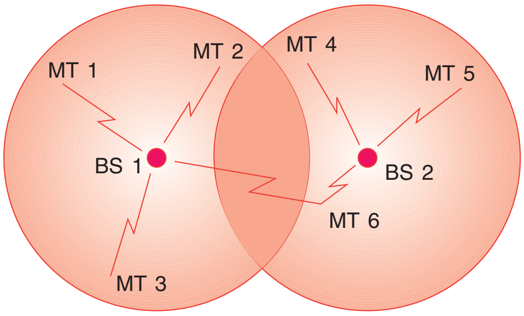


Figure 7 The TS program simulates calls carried by mobile terminals and keep records of the interaction between calls. These interactions are established through radio interference and radio resources assigned to each base station

5.1 Distribution

The TG module requests information from the GEO module about the offroad areas and road networks within the simulation area. The total number of users that generate the offered teletraffic is determined from mobile traffic intensity figures for roads and offroads. The resulting offered teletraffic intensity is determined from distributions set up for each network type or offroad type.

The call set-up requests are then randomly picked from the distribution of the total intensity, and distributed over offroad areas and road networks, according to the weights given by call intensity figures. Likewise, the mobile type, terminal type and service types of the call are randomly picked on the basis of weight figures of the type triplet.

5.2 Call characteristics

For each service type, the call characteristics are configurable, where distribution type and expectation value can be set for interarrival time and holding time. The distribution of the call duration can be fixed, uniform or exponential, depending on the service type.

5.3 Mobility

The importance of the mobility model and the parameterisation of it has been discussed widely, since this is a significant part of the foundation for the TS simulation. The approach has been to design a TG by first using a very simple mobility model, and possibly using a more complex mobility model at a later stage. When trying to replicate the essential characteristics of the mobility in a real-world scenario, it is necessary to obtain a high quality level on input data, which may be hard.

The chosen "simple" mobility model in the TG model might be judged somewhat crude as seen from a road traffic planner's view, but may prove to be adequate as long as exact data is hard to obtain.

Traffic flow figures are in the TG associated with the roads and offroads, describing the number of vehicles flowing at which speed and with which standard deviation. Different mobility patterns are created by varying movement speed and direction, controlled by parameterisation of the mobile type. When a mobile approaches a crossing within a road network, the choice of next road to follow is based on weighting the traffic flow out from the crossing. However, the previous road is weighed down to avoid U-turns.

The mobility of offroad mobiles is determined by an algorithm that can be parameterised to provide both a 'brownian movement' pattern of piecewise linear movement within the area.

To sum up the fundamentals of the mobile trace generation part, one should emphasize the following:

a The model produces only one call per trace.[1] The mobility pattern of idle mobiles are of no interest to the simulation as long as location registration is not simulated.

b The trace may be in either an offroad area or along the roads within a given road network.

c The model is based upon the assumption that the long distance mobility pattern of the individual mobiles is of little interest. Cases where car traffic patterns very much concentrate on cars moving into the city centre or out of it, or moving from one area to another, are hard to describe within this model.

An alternative road traffic model has been designed and is partly implemented, in which the aspects of the last bullet item are catered for.

6 Telecom Simulation

The TS module receives from the TG a set of trace descriptions. These traces define the total offered teletraffic for one simulation. Each trace includes a description of the call (service type, start time, holding time, etc.) as well as a description of how the mobile terminal moves during the period when the call is executed. The TS module simulates how the specified GSM network serves these call requests and produces statistical results such as probability of dropped call, handover activity, radio link quality, etc. for the simulation run.

The TS module models the GSM network simulated. This model includes a description of propagation of radio signal in the geography as well as a model of each BS.

The TS module starts simulation by executing calls as they are issued by the TG program and will during one run have several ongoing calls at the same time. Figure 7 depicts a possible state of the simulator during a run. The TS module models each mobile terminal and base station transmitter and their transmitted power and updates the total interference level in the simulated area. It computes how one active call is interfered by the rest of the active calls.

The TS module models each phase of a GSM telephone call. This includes modelling of cell selection, call set-up, additional assign, active phase, handover and call release. For each call being simulated MOBSIM keeps record of the current interference and noise level as well as received carrier signal level. These variables are used for continuous evaluation of the state of the call. These evalua-

[1] *However, when call retries occur due to e.g. congestion, these retries are allocated to the same trace. This is possible because each trace has a certain safety extension to cope with such situations.*

tions will detect whether the following events have occurred: handover criteria are fulfilled, power should be regulated or the call should be aborted. A call may be aborted for one of the following reasons:

- Bad radio signal strength in the area where the mobile terminal is located
- Bad radio signal quality due to high interference level from other terminals and BS's
- Congestion: There are no available resources at call set-up
- Handover congestion: A mobile terminal which moves into another cell is not allowed to perform handover to this cell due to congestion in this cell.

An aborted call will be re-established guided by specified rules. These rules are described more closely later in this chapter.

MOBSIM is well suited for testing different handover criteria and power regulation algorithms. Several criteria for handover have been implemented and tested within the simulator and have verified the need for a tool like MOBSIM for validation of handover algorithms as well as studies of network implemented by network operators.

6.1 A software model of the Telecom Simulation

The TS program is made up by a set of software entities. These entities are depicted in Figure 8. We now give a short summary of these entities even if they are described more closely later in this chapter. Cells which are included in the simulation scenario are represented by base station objects, and logical channels assigned to a cell are handled by a channel pool object.

Call execution is handled by four different entities:

- A link measurement entity which performs the determination of the link communication quality
- A call entity performing the execution of the call including handover, power control and link quality monitoring
- A user entity which models the behaviour of the user upon such events as call abortion and call congestion
- The terminal entity models the characteristics of the mobile terminal and the mobile entity models how the terminal moves in the (off)road area.

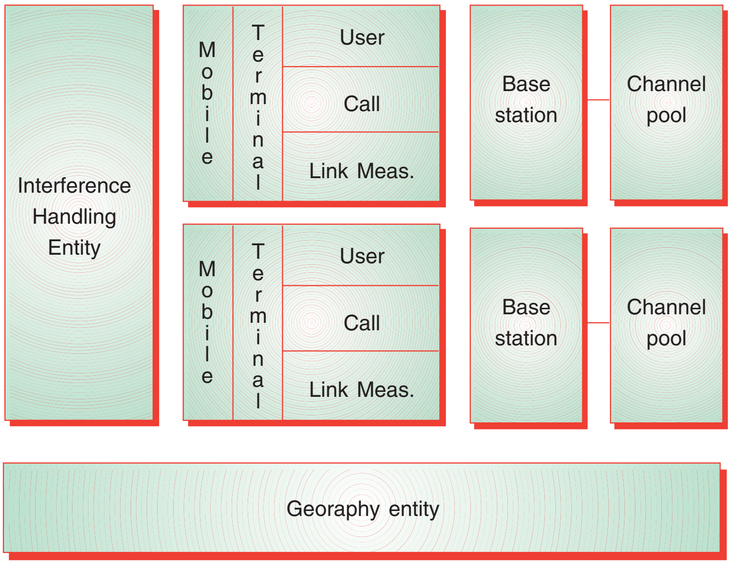


Figure 8 A software model of the TS program

The interference handling entity handles the propagation of radio signals and interference between the radio links. The geography entity models the road networks and path loss values for each individual cell.

6.2 Network modelling

The TS module models the GSM network. This network may form a virtual network with identical hexagons or representation of a real GSM network as it is implemented by a network operator.

The BS entity depicted in Figure 8 is a model of a GSM cell and equipment/ functionality associated with this cell. It is associated with a set of parameters which control the output power of the BS and parameters which control output power of the mobile terminal within the cell. A BS is also modelled with some parameters which describe the transmitter/receiver and antenna equipment (i.e. antenna gain and equipment losses).

The channel pool entity is associated with a set of frequencies which are divided into a set of logical channels. These logical channels are available for communication between a mobile terminal and a BS. Two types of channels are implemented in MOBSIM: TCH/F + SDCCH (TACH/F) and SDCCH + SACCH (TACH/8)[2]. The TACH/8 channel is used for communication in the establishment phase and TACH/F is used under 'normal' user-to-user communication.

The channel pool entity is responsible for administration of the channels dedicated for one cell, i.e. it assigns channels for new calls and keeps count of channels in use and not in use. This entity registers the congestion level of the cell and reject new calls and calls handed over to the cell when all channels within the cell are occupied.

Associated with a base station is radio data which describe the propagation of radio signals transmitted or received by the base station. The data is handled by the geography entity and was described in chapter 4.

6.3 Mobiles and terminal entities

MOBSIM models the mobility of the terminal by an entity termed 'mobile'. The mobile entity receives its trace description from the TG module. The mobile will move the mobile terminal along the trace as time progresses in the simulator, and while doing so, collect the appropriate set of radio data.

MOBSIM models mobile terminal entities with the following parameters: maximum output power, transmitter/receiver and antenna parameters. Also, an additional loss parameter is included. This parameter represents losses in the environment of the terminal antenna, due to attenuation inside a car or a building.

6.4 The radio propagation model

A radio signal transmitted from the base station antenna to a mobile terminal

[2] *See [9] for more description on channel configurations.*

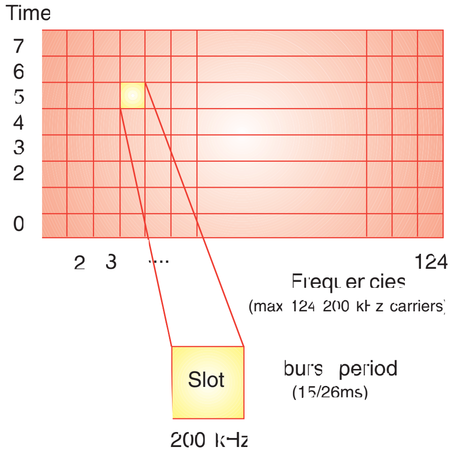


Figure 9 A way of describing the available GSM spectrum as a 2D array. The x-axis denotes the available frequency spectrum, whereas the y-axis denotes the basic time units, i.e. the burst periods

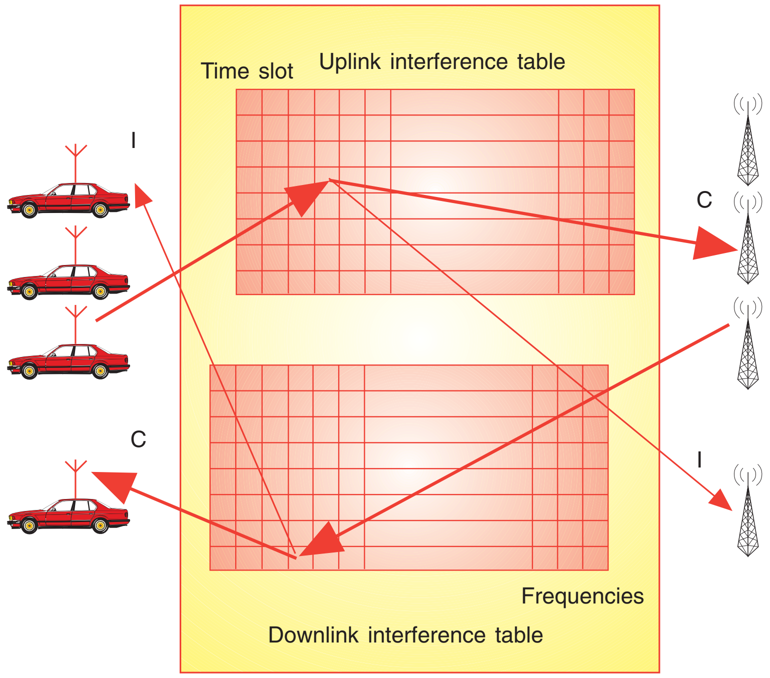


Figure 10 Overall scheme for updating the snapshot of signal and interference levels at spectrum locations

antenna will experience attenuation or path loss. This attenuation is due to free space propagation loss, shadowing and absorption. The uplink attenuation is always assumed to be equal to the downlink attenuation. Several factors may be included in the computation of the path loss. Written material on the subject can be found in [1] and [2].

Usually, a cell planning system (CPS) is used in conjunction with MOBSIM. The CPS produces path loss values for a base station in a geographical 'bin'-spot.

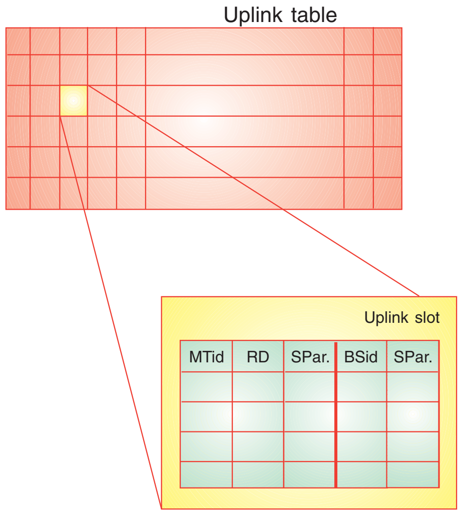


Figure 11 The data structure of the uplink table with its uplink slots. Within each slot representation is given a list of the mobile terminals which are currently sending in the slot together with a list of base stations which are currently listening at that slot

MOBSIM performs the following computation for the received power in the mobile terminal or in the base station receiver.

$$
\begin{aligned}
&\mathit{ReceivedPower} = \\
&\mathit{TransmittedPower} \\
&+ \mathit{transmitterAntennaGain} \\
&- \mathit{transmitterLoss} - \mathit{attenuation} \\
&+ \mathit{receiverAntennaGain} \\
&- \mathit{receiverLoss} \qquad (6.1)
\end{aligned}
$$

6.5 Interference calculation

As outlined in the GSM specifications, the GSM uses a combination of FDM and TDM. For GSM 900, 124 different carrier (centre) frequencies are used, where each carrier occupies a certain portion of the total bandwidth. For the interference calculation module of MOBSIM we use the following subdivision of the spectrum, see Figure 9:

- 124 intervals along the frequency axis, where each interval represents the portion occupied by carriers using a given centre frequency. Adjacent intervals overlap.
- 8 intervals along the time axis. This subdivision is based on the frame structure of 8 time slots, in GSM terminology burst periods. The fact that some signals are not present in every frame is therefore not contained in this subdivision.

Each combination of frequency interval and burst period is referred to as a radio resource. Thus, we have a total of 124 x 8 different radio resources.

Since we may assume that interference does not occur between uplink and downlink (e.g. that no base station transmission may interfere with the signals from mobile stations to other base stations), MOBSIM employs two separate tables for updating the status of interference, see Figures 9 and 10.

The uplink interference table comprises all information necessary to get a clear picture of how any base station that listens to any mobile terminal at any specific slot receives this mobile's signal and also how much interference the base station receives from other far away mobiles currently transmitting in the same slot. The data structure is depicted within the 'Uplink Slot' in Figure 11. Within each 'Uplink Slot' there are two lists present: One for the mobile terminals that are currently transmitting within the slot, and one for the base stations that are currently listening within the slot. Each entry in the first list contains a) an identifier of the mobile terminal (MTid), b) the set of radiodata that are valid for this mobile terminal, and c) all necessary characteristics of the transmission of the mobile terminal (SPar, e.g. on which logical channel the mobile is currently transmitting and with which output power).

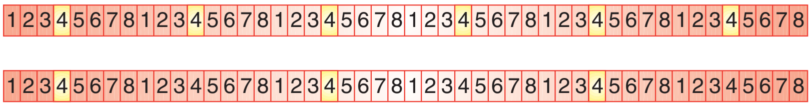


Figure 12 At the top, a receiving party is listening to its corresponding transmitting party at full rate speech channel in time slot No. 4. Below, an interfering party is transmitting in time slot No. 4 on a half rate speech channel

Each entry of the second list contains a) an identifier of the base station that is currently receiving information (BSid), and b) all necessary characteristics of the reception of the base station (SPar, e.g. on which logical channel the mobile is currently receiving). In Figure 11, these two lists are represented by a table in which the two entries for a) the transmitting mobile and b) the base station to which the mobile is sending its data are located on the same row.

The aspects of how interference is dependent on the logical channels used may deserve a comment. A receiver receives at a certain time slot either in all TDMA frames or a subset of the TDMA frames in a multiframe, dependent on the logical channel that it employs. A transmitter does therefore not necessarily interfere with a specific receiver for every burst it transmits. When calculating the impact that a certain transmitter may have on receiver listening within the same time slot but possibly at another logical channel pattern, one has to consider the transmission and reception ‘overlap’ between the parties as given by the logical channels that they currently employ. An example is depicted in Figure 12, where an interferer on a half rate speech channel transmits on the same time slot as a receiving party employing a full rate speech channel. In MOBSIM, the influence of an interferer within the same time slot but transmitting on a different logical channel is averaged with respect to how many times per multiframe the transmitting of the interferer coincides with the receiving of the party being interfered. In the example depicted in Figure 12, the influence of the interferer is 1/2 of what it would have been if the transmitting and the receiving party had employed the same logical channel. Note that if the upper party of Figure 12 was the transmitting one and the lower party was the receiving one, the influence of the interferer would have been 1/1 of what it would have been if the transmitting and the receiving party had employed the same logical channel.

As may be seen from e.g. Figure 12, the interference calculations are based upon the assumption of full synchronism, i.e. synchronism at the levels of time slot, TDMA frame and multiframe. For BTSs under the same BSC, one may consider this to be an appropriate assumption. However, between different BSC areas the assumption may not be valid.

Some final comments should follow the description of the interference calculation.

One concerns the DTX feasibility of GSM. MOBSIM is prepared for handling networks employing the DTX feature. A ‘speech burst pattern’ is defined within the input data, and according to this distribution, MOBSIM randomly picks speech burst lengths. The interference calculation module deals in detail with these speech bursts when producing interference figures.

Another comment concerns transmission on beacon frequencies. MOBSIM fully complies with the GSM specifications in the respect that beacon frequencies are permanently regarded as interferers, independent of the actual signalling or teletraffic that is carried by them.

Finally, it should be said that frequency hopping is not yet dealt with by MOBSIM. However, the interference calculation module is designed in such a way that frequency hopping may be introduced without severe problems.

6.6 The link measurement model

The link measurement entity depicted in Figure 8 is responsible for two things: Evaluation of the probability for successful layer one frame transmission and for producing measurement figures for the link quality as is done in the real GSM system.

The interference calculation in the previous subchapter produces carrier signal (C), co-channel interferers (I_{co}) and adjacent channel interferers (I_{ad}) levels for receivers in the simulator. These signal levels are used to compute a probability for successful transmission of one symbol over the radio channel.

There are extensive coverage in literature ([4] and [5]) on how symbol error rate is changing according to changes in different signal components and for different modulation schemes. [6] gives documentation of the relationship between (C, I_{co}, I_{ad}) figures and symbol error probability (termed raw bit error rate) in the GSM system.

Information in GSM is transmitted as frames. These frames are coded with a FIRE code and then with a convolutional code. These codes are used to detect transmission errors in received frames and correct them. [7] gives documenta-

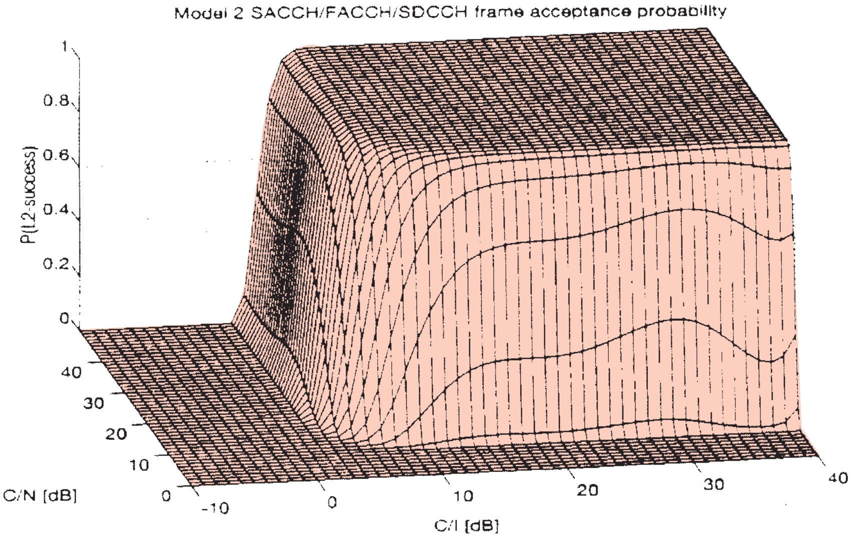


Figure 13 Probability of success of an L2-frame transmission as a function of C/N and C/I

tion on how the frame transmission success probability is a function of the raw bit error rate. Figure 13 shows the result given in [7] for the dependencies of frame transmission success probability upon the (C/I, C/N) figures.

The mobile terminal receiver and the base station receiver in the GSM system performs measurements of received signal level and detected bit error rate (RxQual) over a measurement period (480 ms). In MOBSIM received signal level is calculated by equation 6.1 of section 6.4. The RxQual value is simulated in [6] as a function of (C/I,C/N).

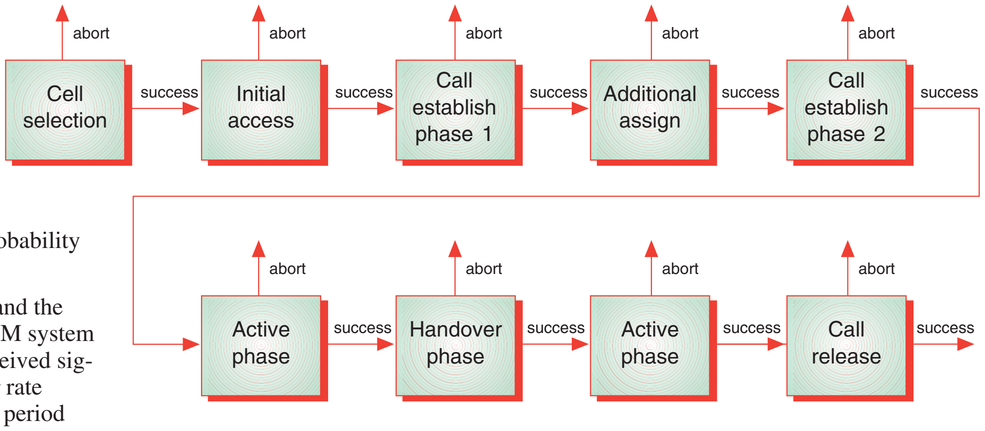


Figure 14 The stages of the call execution model

6.7 The user model

The TS program receives a description of a call from the TG program. This description contains the following items: The point in time when the call should be established, holding time for the call and a description of the behaviour of the user performing the call.

The user behaviour description includes how the user reacts upon call congestion and how he reacts upon abortion of the call. The description also describes the burstiness of the user i.e. when he talks and when he does not talk. This description is for simulation of DTX.

6.8 The call execution model

MOBSIM divides simulation of a call into stages. These stages are depicted in Figure 14. The stages are executed in a sequential order and abortion/failure in one of the stages will abort the call. This call abortion will be handled in accordance with the user model of the call.

The cell selection stage is responsible for performing the cell selection algorithm as described in the GSM standard. The selected base station is the one with the highest received power, taking into consideration path losses as described in equation 6.1 of section 6.4. For the chosen base station the GSM Cx > 1 criterion is checked. C1 is defined in [3] with a logarithmic formula with the requirement C1 > 0. The equivalent non-logarithmic formula is defined as Cx:

- Cx = A /max(B,1)
- A = Received level / RXLEV_ACCESS_MIN
- B = MS_TXPWR_MAX_CCH / Maximum RF output power of the MT.

If Cx ≤ 1 then the cell selection has failed. The RXLEV_ACCESS_MIN and MS_TXPWR_MAX_CCH parameters are system parameters specified for each cell. Upon abortion of cell selection the user model will handle this in accordance with the user behaviour rules.

The initial access stages which are performed after a successful cell selection may be divided into three substages: Access on random access channel (RACH), assignment of a logical channel and communication on the access grant channel (AGCH). MOBSIM performs no simulation of the random access and access grant channel communication. This means that these stages are presumed perfect and they will therefore never fail. The assignment of a logical channel is, however, performed. This assignment is successful if there are available channels in the cell. If all logical channels within the selected cell are in use then the call is aborted and the simulator will handle this abortion in accordance with the user behaviour rules.

The call establishment stage is performed after a successful initial access stage. This stage includes transmission of a set of messages between the mobile terminal and base station over the channel assigned in the initial access phase. These signalling messages are sent on OSI layer 3. Layer 3 messages consist of a number of octets, and they are directly packed into L2-blocks. If more than one L2-block is necessary, the message is segmented at the transmitting side and re-assembled upon reception. Completion of layer 3 messages are dependent on retransmission of L2-blocks. Layer 2 messages are retransmitted, if they are corrupted, according to the LAPDm protocol. [8] contains a set of equations which determine the relation between probability for a successful layer 3 message and probability for successful frame transmission depicted in Figure 13. These functions are used in the call establishment stage to determine whether the call establishment is successfully performed or not; C/I and C/N figures are necessary since this determination is produced by the link measurement entity.

The additional assign stage models the procedure in the GSM system where a call changes its channel type from TACH/8 to TACH/F. This stage may be divided into two subparts:

a The determination as to which cell the additional assign procedure should be performed, and allocation of a TACH/F channel on this cell

b Execution of handover signalling.

Substage a is dependent upon the algorithm employed; however, if there are no free channels the call will be aborted. Substage b consists of a set of messages transmitted on the old and new channel. The modelling of this stage is equivalent to modelling of the call set-up stage.

The call establishment stage after additional assign models the transmission of signalling messages for completion of the call establishment. This is modelled equivalent to the other signalling stages. The release stage is modelled in the same way.

The call entity performs three activities in active state: handover evaluation, power control and link monitoring. We will not cover the handover or power algorithm because these will be different from network to network. However, decisions about handover to other cells (or to a channel within the cell, i.e. intra handover) and power regulation will be based on the RxQual and RxLev figure produced by the link measurement entity.

The link monitoring activity is responsible for controlling the quality on the radio link. Both uplink and downlink qualities are estimated through a counter which detects the rate of losses of SACCH frames. This counter is set to an initial value at start up of the call or after a successful handover. The mobile terminal will increment the counter by two if a SACCH frame is successfully decoded, otherwise it will decrement the counter by one. The call is aborted when the counter reaches zero. The counter will never exceed its initial value. A possible scenario is depicted in Figure 15.

7 Logging and statistics

7.1 The process of collecting, merging and sorting statistical data

MOBSIM produces two main types of output data:

- Log records
- Statistical data in terms of produced values of certain estimators.

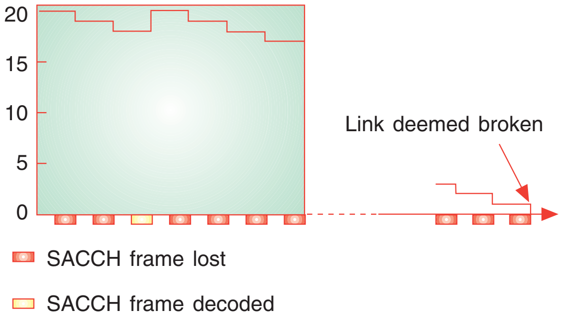


Figure 15 The Radio Link Monitoring algorithm

Both types rely on two concepts that are crucial in MOBSIM: Significant Event (SE) and Processed Event (PE). A Significant Event is a data sample that describes a specific, pre-defined event being such that no prior data collection has been necessary. A Processed Event resembles a Significant Event, but for establishing a Processed Event, prior data collection has been necessary. In general, a Processed Event may be established when a certain set of Significant Events

Table 1 Stochastic variables that may be measured within MOBSIM

Stochastic variable that may be measured	Short hand	Comment
Fraction of time of a call for which $C/I < \Phi_{C/I}$	CI	Gives an indication of the average quality of the calls that are performed with respect to interference. The measure requires definition of the threshold value
C/I just before a handover	CI_B	Should be seen in conjunction with the next variable, as the two together give an indication of the average improvement of a handover with respect to interference
C/I just after a handover	CI_A	As above
Call congestion for traffic channels per base station	CC_T	General capacity measure
Call congestion for control channels per base station	CC_S	General capacity measure
Time congestion for traffic channels per base station	TC_T	General capacity measure
Time congestion for control channels per base station	TC_S	General capacity measure
Time between dropped calls	TD	Contributes to the relative dropped call rate, which is a very essential measure within the scope of quality of service
Time between handovers	TH	Contributes to the relative handover rate, which is a very essential measure when detecting network configurations that produce affluent handovers
Time in high load state for traffic channels	TL_T	Gives an indication of the average traffic channel load. The measure requires definition of the threshold value
Time in high load state for control channels	TL_C	Gives an indication of the average control channel load. The measure requires definition of the threshold value
Duration of successful inter handovers	THO	Gives an indication of whether of not there are transmission problems preventing the handover signalling procedures to work as smoothly as they can
Probability of successful call set-up and connection phase	PC1	Gives a measure of the rate of 'no problem' calls, i.e. calls that are neither congested nor disconnected
Probability of successful call set-up without interruption	PCE	Gives a measure of the rate of calls in which the connection is established (but which may later be disconnected)
C/I during the call period	CI_C	Gives a measure of the average level of C/I for a call (averaging is taken both over the call period and over the call samples)

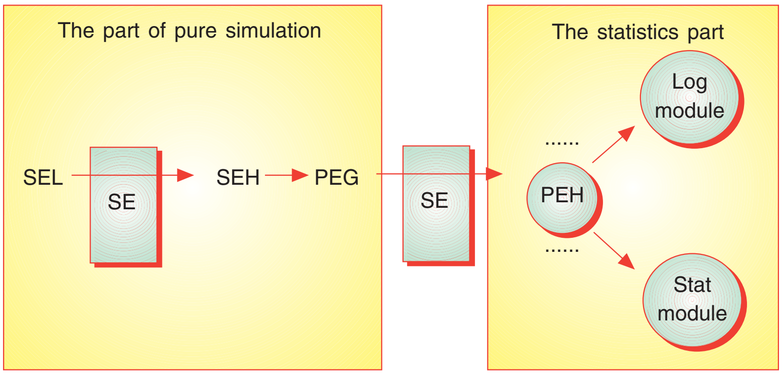


Figure 16 The conveyance of relevant information due for log or stat modules from the simulator

has been obtained. Processed Events are often used to characterise certain features of calls, for each of which a specific set of Significant Events (e.g. call set-up, handover, call release, forced or orderly) needs to be collected.

The two main software modules concerning logging and statistics are

- the part of pure simulation, i.e. the part of the simulator that genuinely describes the mobility and teletraffic handling of the system
- the statistics part, i.e. the part of the simulator that only deals with the processing of the data collected for logging or statistical purposes.

Within the pure simulation part, the data flow starts at the software locations where the significant events are detected and described (SEL). At these points, significant events are generated and sent to the appropriate significant event handler (SEH). The significant event handler takes care of the arriving SE and routes it to the appropriate processed event generator (PEG). The PEG builds up a processed event (PE), and transfers it to the processed event handler control (PEHC) within the statistics part. PEHC routes it to the appropriate processed event handler (PEH), which in its turn conveys it further to either the log or the statistics module.

The chain SEL → SEH → PEG → PEH reflects the process of collecting, merging and sorting statistical data ever more general when approaching the statistics part.

7.2 The stochastic variables to be measured

The stat module offers to the user of MOBSIM a set of stochastic variables that may be measured, as presented in Table 1.

7.3 Control of run time length based on confidence control

As for most time series simulations, the statistical data that is requested may be given together with an accuracy interval and confidence level for the tests of checking whether or not that accuracy is reached. The method employed in MOBSIM of currently checking whether or not the stated criteria are reached, is based upon the techniques of putting the observations in batches over a certain size, and then to perform variance and correlation tests on the sequence of these batches. The method is described in detail in [10].

8 Conclusion

MOBSIM is a powerful tool for the radio planning department of a GSM operator that has service quality as a competitive objective. Its merits are first of all the capability of verifying site planning and frequency allocation schemes with respect to teletraffic carrying capacities, but it also opens for analysis of system parameters and functionality like the handover or power control algorithm to be applied. It may also serve the purpose of being a tool for a more extensive exploration of problems in the field of mobile communication, either as it is or with minor extensions.

9 References

1 Parsons, D. *The mobile radio propagation channel.* London, Pentech Press Publishers, 1994. ISBN 0-7273-1316-9.

2 Lee, W C Y. *Mobile communications engineering.* New York, McGraw-Hill, 1982. ISBN 0-07-037039-7.

3 ETSI. *GSM Recommendation 05.08 Radio Sub-System Link Control.*

4 Proakis, J G. *Digital communications, 2nd edition.* New York, MacGraw-Hill, 1989. ISBN 0-07-050937-9.

5 Van Trees, H L. *Detection, estimation and modulation theory, part 1.* New York, Wiley, 1968. ISBN 0-471-89955-0.

6 Lehne, P H et. al. *Mobilkommunikasjon fleksibel aksess : radiokanalen.* Kjeller, Norwegian Telecom Research, 1993. (TF-report R 40/93.) (In Norwegian.)

7 Lehne, P H. *Tabulating the GSM layer 2 frame success probability.* Kjeller, Norwegian Telecom Research, 1994. (MOBSIM project report TF_PHL-Report-4/94.)

8 Lehne, P H, Svebak, O D. *Event success probabilities based on radio channel quality.* Kjeller, Norwegian Telecom Research, 1994. (MOBSIM project report TF-PHL-report-1/94.)

9 Mouly, M, Pautet, M-B. *The GSM system for mobile communications.* Palaiseau, published by the authors, 1992. ISBN 2-9507190-0-7.

10 Berg, T. *Estimering av middelverdi og konfidensgrenser for en stasjonær korrelert tidsserie.* Kjeller, Forsvarets Forskningsinstitutt, 1991. (FFI/Notat-91/7038.) (In Norwegian.)

Train radio system for Norwegian State Railways

BY BJØRN OLAV SOLBERG

1 Introduction

SCANET is the train radio system for Norwegian State Railways. The radio system is developed by Ascom Radiocom in cooperation with Norwegian State Railways. The installation of the system started in 1993 and all the main lines will be covered by the end of 1996. The main purpose of the train radio is to permit radio communications between the engine driver and the train dispatcher (see Figure 1) and other subscribers involved in railway operations. The main parts of the system are the Mobile Station located in the engine, Base Stations along the railway lines, and the Central Traffic Control. The system is connected to the internal railway telephone network and acts like a normal telephone. The Automatic Train Control (ATC) is connected to the train radio. ATC is based on information received from 'passive beacons' located between the rails. One of the main purposes is to determine the train position. The position data is transferred via the train radio into the fixed network and ends up at the train dispatcher. The train radio will contribute to a safer and more efficient train transportation.

2 System description

2.1 General

The SCANET train radio is a system for supporting train transfer in Norway. The main purposes are voice communication between the engine driver and the Central Traffic Control and data communication (see Figure 1). But it is also possible to communicate on three other radio networks in the Norwegian State Railways.

The train radio system operates in the UHF frequency band, 457.600 – 458.475 MHz (up-link) and 467.600 – 468.475 MHz (down-link), all the networks included. Frequency Division Duplex (FDD) is used to separate the two transmission directions with a duplex spacing of 10 MHz. Frequency Division Multiplex (FDM) is used with a channel spacing of 25 kHz giving 34 full duplex channels for all the radio services. The total number of channels for the train radio communication is 12 duplex channels and the modulation type is PM for voice and FM for data. The power output for the Base Station and the Mobile Station is 10 W.

All main railway lines in Norway are divided into different radio areas identified by a specific number. Each area is connected to a Central Traffic Control and there are several Base Stations within one area. Each area is allocated a group of 4 channels. The Base Station operates like a repeater station between the Central Traffic Control and the Mobile Station. Messages are transmitted in both directions simultaneously. All areas are supervised and monitored by the Dispatch Centre. Speech and data signals are transmitted simultaneously using speech compression and time division multiplexing. There is only one speech connection at the same time in one area. There is also a connection to the internal railway telephone system.

A call from a train to the train dispatcher is shown on a screen in the Central Traffic Control. The screen indicates the train number and position. The train dispatcher chooses whom to answer first. A call from a train will also be indicated if the train dispatcher is talking to another train. The train number and position is continuously indicated on the screen at the Central Traffic Control.

As we can see from Figure 2, the Mobile Station is the centre of the complete communication network. All communications are carried out to and via the Mobile Station. The radio station is divided into two radio transceivers and one scanner. Train Radio Communication (TRC) and Operation Radio (OPR) are connected to one transceiver, and Station Radio (STR) and Internal Radio (IR) are connected to the other transceiver. The scanner automatically updates the best TRC Base Stations (or OPR base stations).

The importance of man-machine interface in the system is taken into account for the entire system. The users have been involved during the whole project period to form the best user interface.

2.2 Network construction

The radio system can communicate on four networks (see Figure 2):

- TRC network (Train Radio Communication) – central traffic control and engine drivers
- OPR network (Operation Radio) – construction and maintenance services
- STR network (Station Radio) – railway station and shunting personnel
- IR network (Internal Radio) – personnel on board the train.

The TRC network is a duplex transmission of speech and data between the Mobile Station in the engine and the Central Traffic Control. Speech and data signals are transmitted simultaneously using time division multiplexing. The data are coming from the TRC system itself and the Automatic Train Control (ATC). The engine driver also has direct access to the railway telephone network via a PABX (Private Automatic Branch Exchange) connection at the Central Traffic Control.

The OPR network is a speech service radio network using semiduplex transmission. It is used for maintenance service along the railway line. The network consists of handheld radios which can communicate with base stations or other handheld radios. The transmission between the TRC Mobile Station and an OPR base station is duplex. If trouble occurs in the TRC network, the OPR network can be used by the engine drivers as a stand-by system. It is also connected to the PABX.

The STR network is located in the station and shunting area and the system operates on semiduplex transmission to the handheld radio and on duplex to the base station.

Communication within the train is possible in the IR network. The handheld radios are mainly used by the train guards on a moving train. The transmission between two handhelds is simplex and semiduplex with Push-To-Talk (PTT) to the Mobile Station and the TRC Base Station. The handheld can also have direct access to the railway telephone network. It requires two channels, so that two crossing trains do not disturb each other.

2.3 Central Traffic Control

The Central Traffic Control consists of one control cabinet and several operator consoles. The control cabinet contains the Front Ends and the Dispatch Centre (see Figure 3).

The Dispatch Centre consists of a central processor and base band switching arrangements. The switch executes the interconnection of speech and data signals from the dispatcher consoles to the Front Ends and the assigned area. It is possible to connect 8 operator consoles and 8 Front Ends to the Dispatch Centre. One operator console can handle up to 6 Front Ends. Several computers (for service purposes) and a printer can also be connected.

The Front End is the interface between the Dispatch Centre and the area lines.

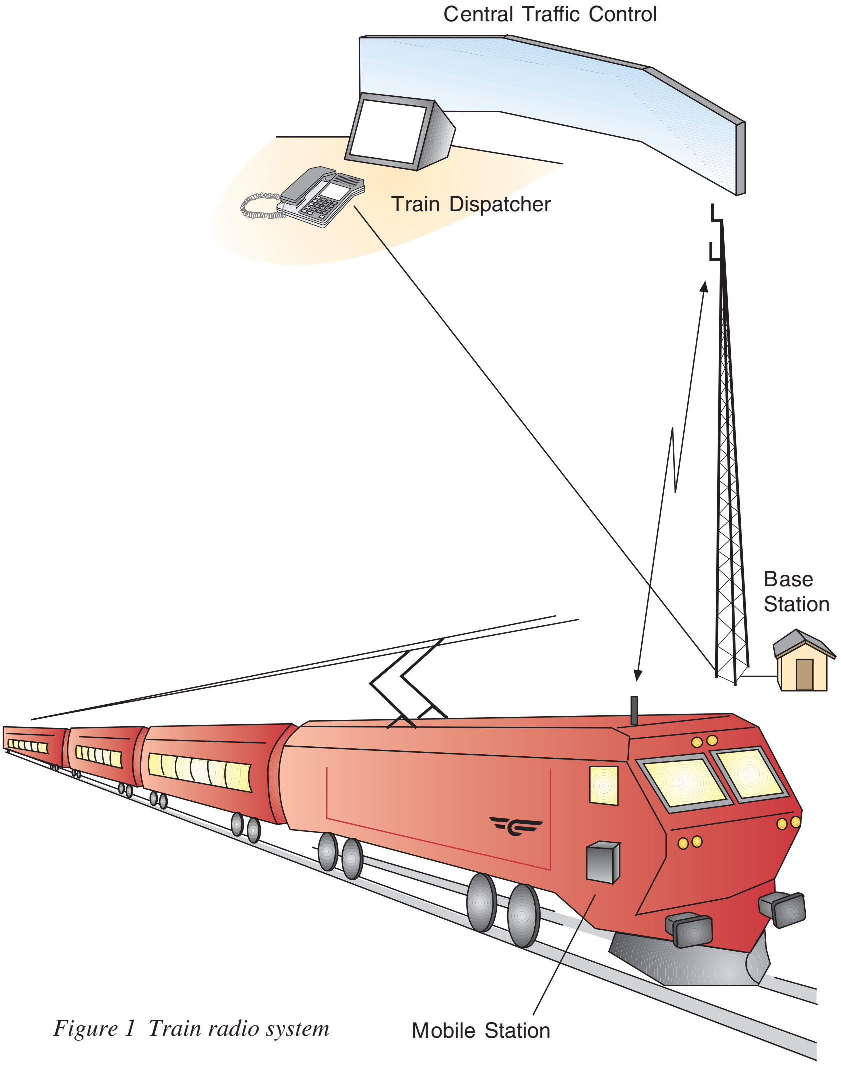


Figure 1 Train radio system

The Front Ends control all the Base Stations and provides synchronisation of speech and data transmission. One Front End can handle up to 30 Base Stations. Each Front End has an interface to the telephone network for speech calls between train subscribers and telephone subscribers. The access to the PABX is still working if the Dispatch Centre fails. Speech compression and expansion takes place in the Front End and in the Mobile Station.

The operator consoles are located at the operation desk at the Central Traffic Control and they are used by the train dispatchers (see Figure 4).

2.4 Base Station

The Base Stations are located approximately every 7 km along the railway line (see Figure 5). A four-wire line (area line) links the parallel connected Base Stations to the Front End in the Central Traffic Control (see Figure 2). For radio areas far away from the Central Traffic Control, the communication goes via fibre optics links on fixed 64 kbit/s or 2 Mbit/s channels. The connection to the four-wire line is done facing a high impedance (approximately 10 kOhm). Up to 30 Base Stations can be connected to one radio area (Front End). The Base Station consists of two parts, a radio part and a data part. The transmitter is continuously in operation and transmits a polling signal periodically which is used by the Mobile Station for a login procedure. It also continuously transmits an identification code specific to its own area. This means that a Base Station does not effect Mobile Stations outside its area.

The Base Station transmits speech signals in compressed form simultaneously in both directions together with slotted data using Time Division Multiplexing (TDM). To achieve synchronisation between speech and data blocks, all the Base Stations must be correspondingly delay adjusted in respect to the Base Station which has the longest communication link to the Dispatch Centre. The Base Station located nearest the Dispatch Centre will have maximum delay.

The data sent to the Base Station are not passed through transparently. They are decoded by modems in store and forward mode before retransmitting. The data is transmitted from the Dispatch Centre to the Base Station and from the Base Station to the Mobile Station and the other way round in two separate polling cycles.

Figure 6 indicates a typical Base Station installation.

2.5 Mobile Station

The Mobile Station is installed in the engine (see Figure 7) and the control unit is located at the control panel in the driver's cabin (see Figure 8). The Mobile Station has two independent transceiver sections and two antenna switch connections. The main transceiver operates in the TRC/OPR network and the second transceiver operates in the STR/IR network (see Figure 2). In addition, the receiver section operating in the TRC/OPR network is equipped with a scanner which carries out a procedure for selecting the correct Base Station within the area. The scanner continuously searches for the Base Station with the highest field strength level. Before the receiver is tuned to a new stronger channel, the signal level of the new channel must exceed a level threshold compared to the old one (hysteresis). The Mobile Station switches to the new channel automatically. The scanner has a minimum reference level and if the signal is above this level, the channel will not be changed, even if a stronger signal occurs. The system provides handover between the Base Stations. The handover to another Base Station during a speech connection is almost unnoticeable to the user.

A call can be set up as soon as the engine driver has completed the login procedure. The engine driver registers his area and train number. Within a particular area, the scanner selects any of the four available channels which provides the best link to a Base Station. If an area change takes place during a call, a warning tone is transmitted from the Mobile Station. The communication is then disconnected and the area change is carried out automatically. The area change is trigged by a 'passive beacon' located between the railway lines (see Figure 9). As soon as the area change has been completed, a

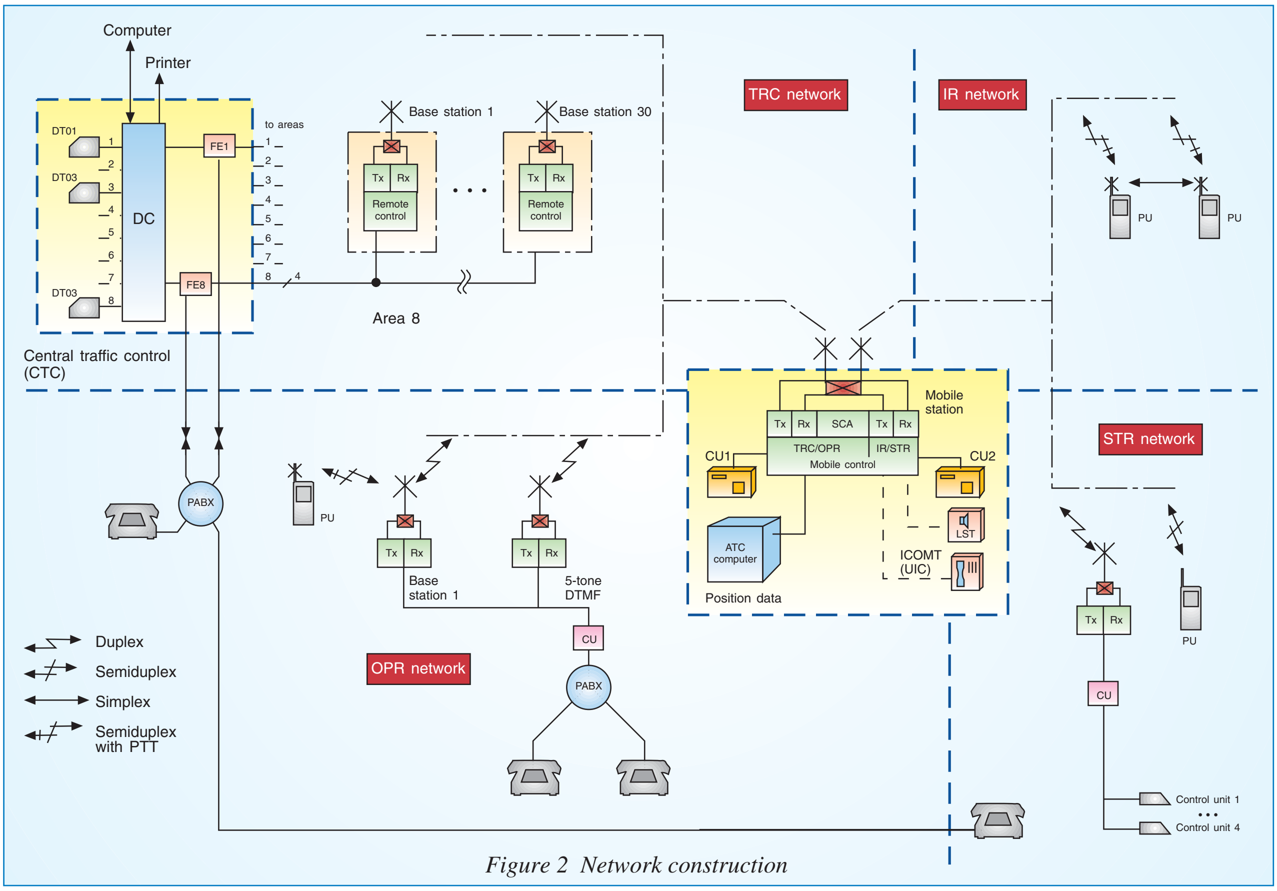


Figure 2 Network construction

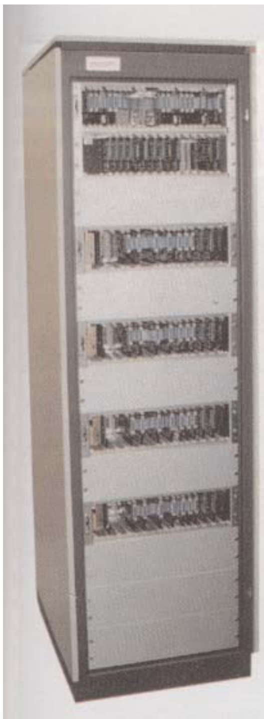


Figure 3 Dispatch Centre (upper part) and four Front Ends (middle part)

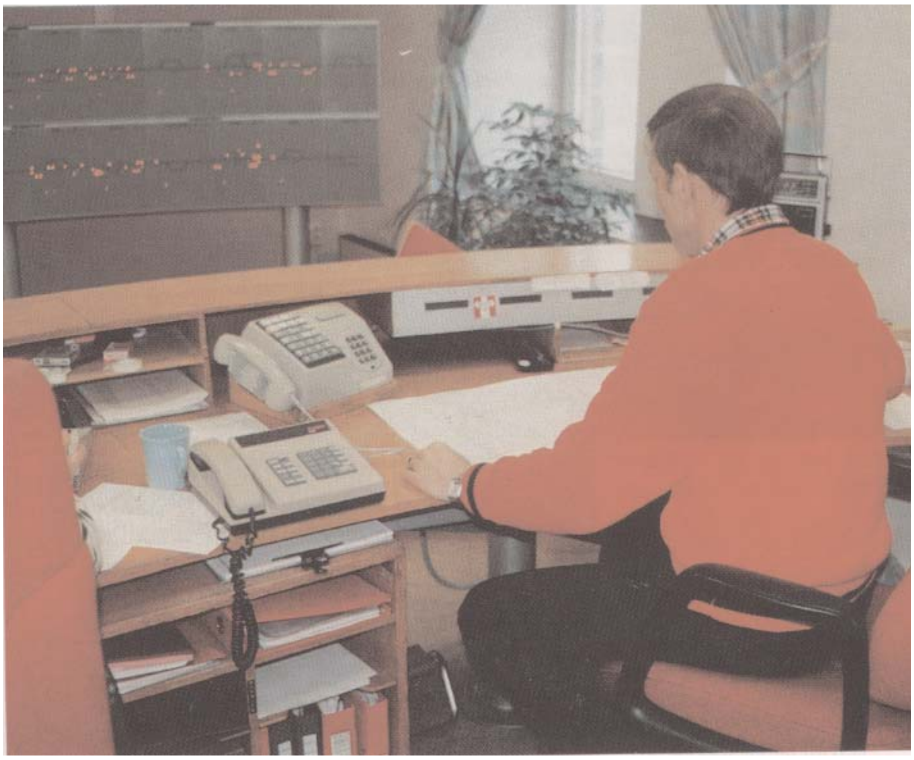


Figure 4 The operator console located at the Central Traffic Control (beside the blue cup)

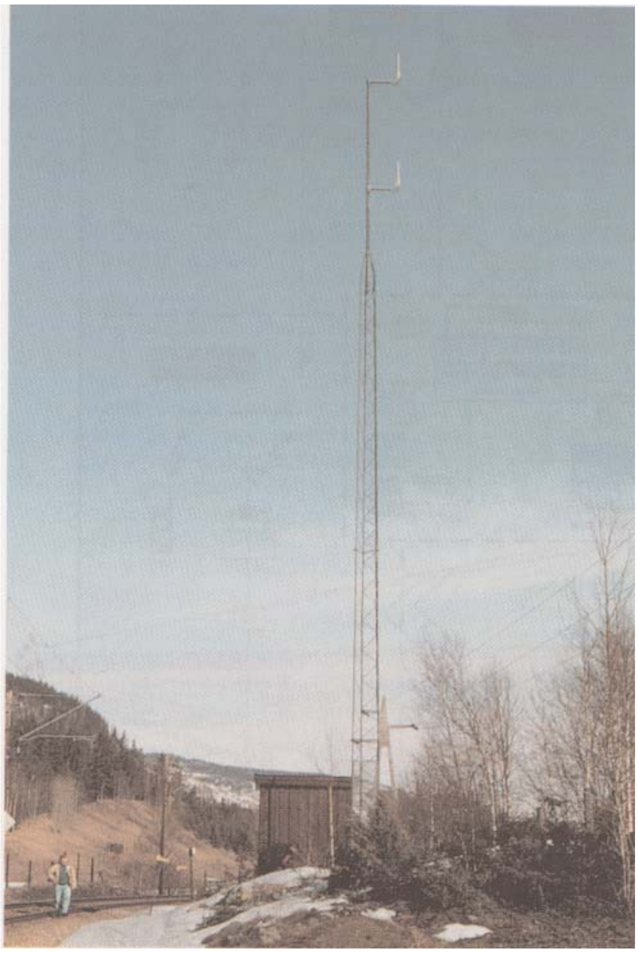


Figure 5 Typical Base Station

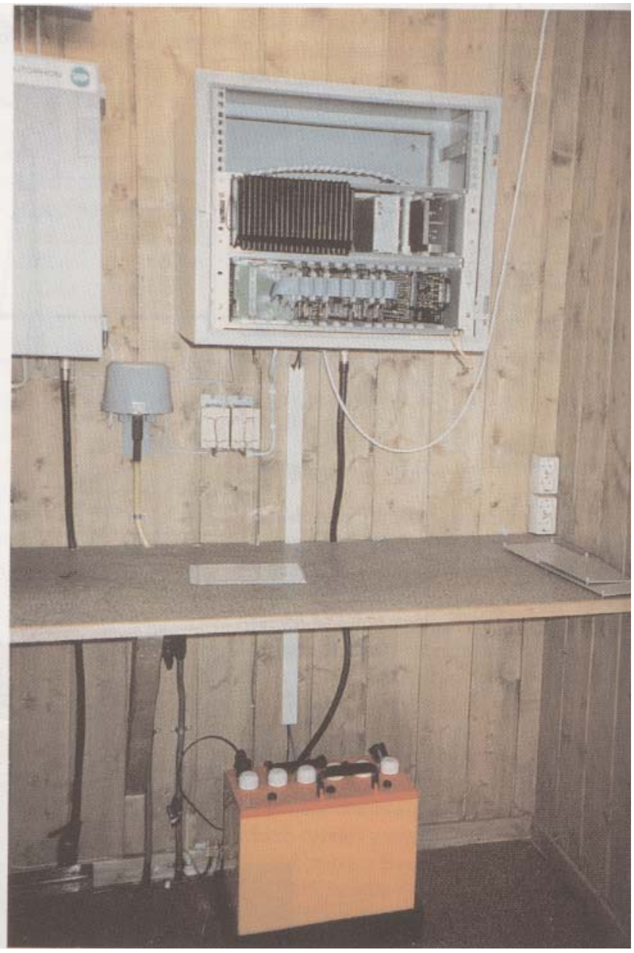


Figure 6 Typical Base Staton installation

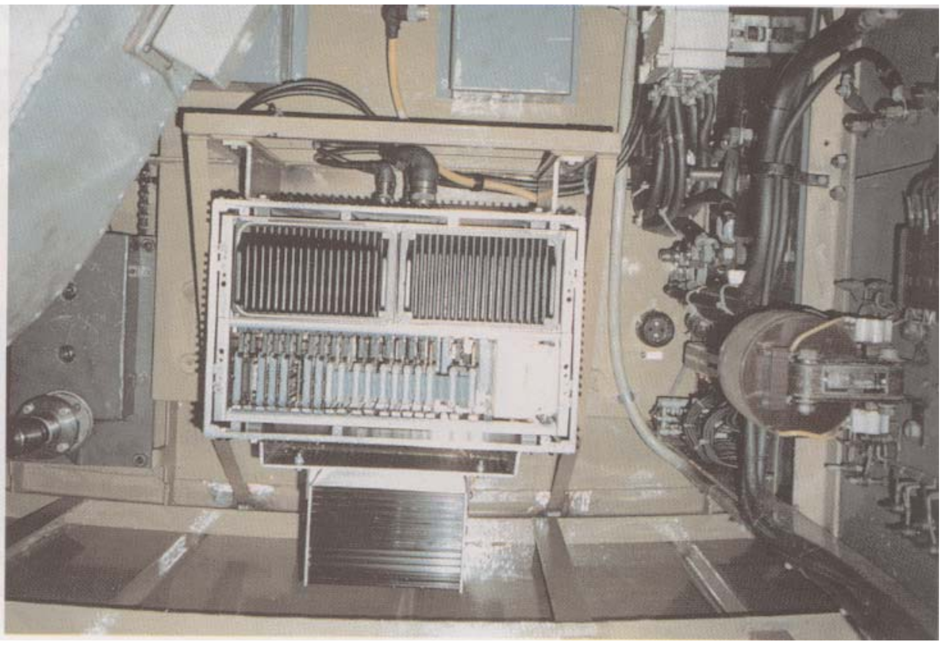


Figure 7 Mobile Station installed in a train engine

call can be set up again. In most cases, the changeover is carried out within the premises of a railway station. If the automatic changeover is not trigged, the engine driver has to enter the new area number manually after resetting the ATC because the ATC data has the highest priority.

The Mobile Station can be connected to the inter-communication system and the engine driver can use the internal loudspeakers in the train.

3 Speech and data transmission

3.1 Speech compression and expansion

Speech compression is implemented to make the system able to have

- calls from other trains to the Central Traffic Control during a conversation
- ATC signalling number transmission during a conversation
- data from other engines during a conversation
- emergency calls.

The system operates on a four-wire line. Speech and data are transmitted simultaneously thanks to speech compression. If no conversation is taking place on the lines, the data transmission capacity is 1200 bit/s. During a speech connection the capacity is reduced to 1200 bit/s in 260 ms per 1040 ms time slot, giving a total capacity of 300 bit/s. It is possible to upgrade the system with another four-wire line exclusively for data (1200 bit/s). The total data capacity is then 2400 bit/s.

Because of the speech compression, time slots are made available between the speech blocks. The data packets are inserted into these slots using TDM (Time Division Multiplexing). Speech compression and expansion is carried out in the Mobile Station and in the Front End. The free time slots are used to transmit data with a 1200 bit/s modem (see Figure 10). First the analogue speech is going through an analogue/digital converter. The speech compression is achieved by reading in the sampled data with a low data rate for a period of 1040 ms and writing it out at a faster data rate. All the frequency components in the base band are multiplied by 4/3. From Figure 11 we can see that the upper part of the base band is cut off by the radio link. The 1040 ms speech slot is therefore reduced to 780 ms giving a compression rate of 0.75.

Expansion uses a similar procedure in the reverse sense. The received sampled data is read at the high data rate for a period of 780 ms and written out at the low data rate for a period of 1040 ms.

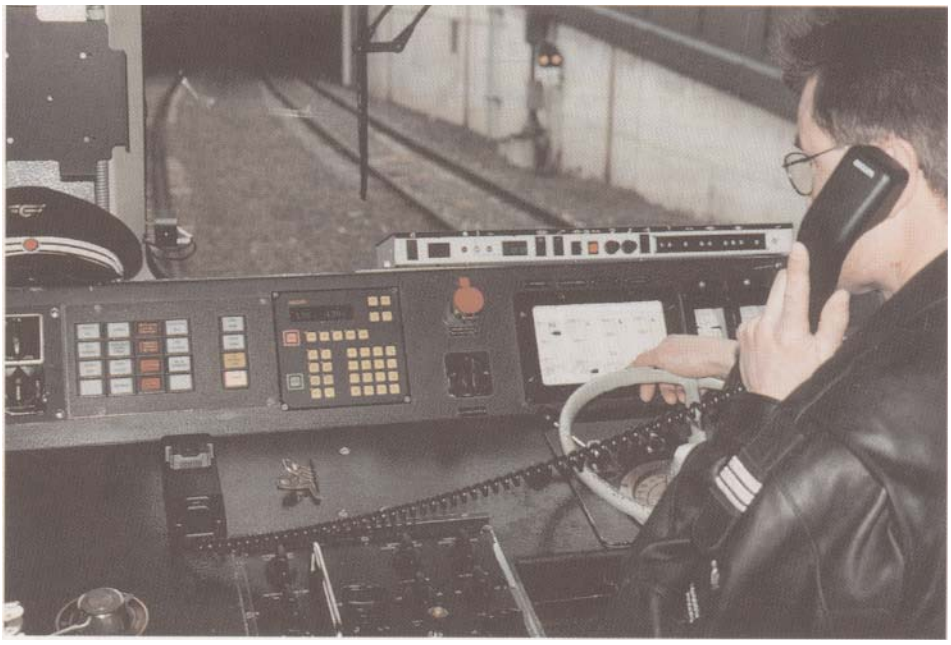


Figure 8 The Mobile Station control unit located in a train engine

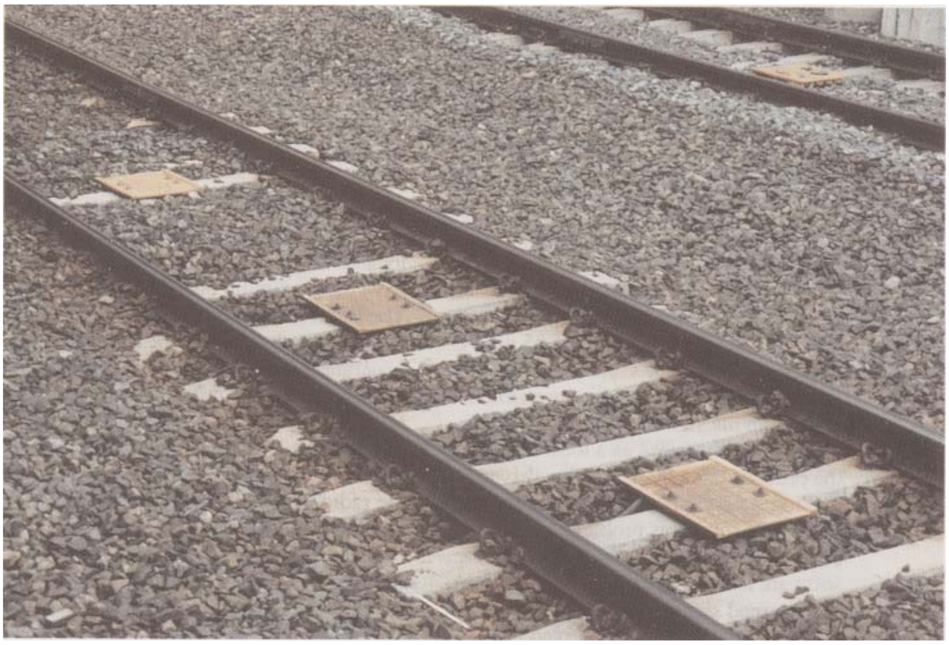


Figure 9 Balise (information points) located between the rails

3.2 Echo cancelling

Speech communication from the engine driver can take place from the Mobile Station, through the Base Station and the Front End ending at a subscriber telephone connected to the telephone exchange. The speech signal travels along a four-wire connection to the output of the telephone exchange where it is normally converted to a two-wire con-

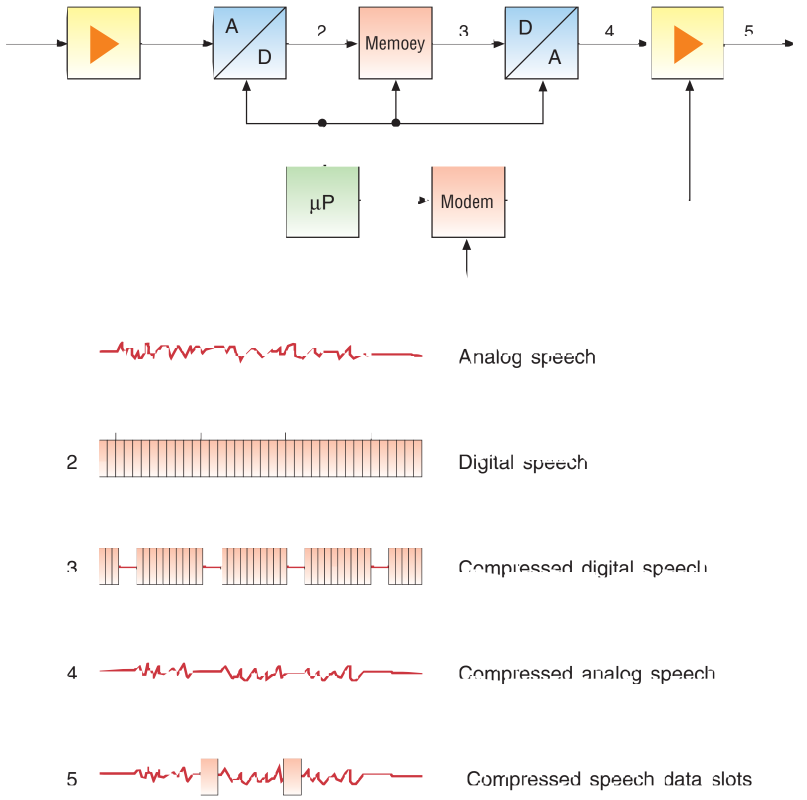


Figure 10 Speech compression

nection for a subscriber. The received signal is attenuated approx. 20 dB in the two-to-four line hybrid converter before it returns along the transmit path back to the engine driver. In the front end the signal is amplified in the order of 15 dB by the AGC (Automatic Gain Control). However, due to speech compression and expansion the signal is delayed by 260 ms, and will be heard as a loud echo in the order of 0.6 s by the engine driver, making it very difficult to hold a conversation. To solve this problem an echo cancelling in the digital signal processor software is introduced which attenuates the returned signal path depending on the magnitude of the signal received from the engine driver (see Figure 12). The echo cancelling is fully digital.

3.2 Synchronisation

Speech and data slots must be synchronised to all Base Stations in an area (Front End). The main goal of the synchronisation is to separate speech and data on the same channel. To achieve synchronisation between speech and data blocks, all the Base Stations must be correspondingly delay adjusted in respect to the Base Station with the longest physical distance from the Dispatch Centre. The delay adjustment is of course zero for this Base Station. The Base Station located nearest the Dispatch Centre will have the maximum delay adjustment. The Front Ends and the Base Stations are equipped with a timer. The Front End generates the master time and periodically sends synchronisation signals to all Base Stations, approximately every 2 minutes. The Base Stations transmit the synchronisation signals periodically over the radio link to the Mobile Station, approximately every 3 seconds (see Figure 13). The synchronisation over the radio link is only necessary during speech mode.

3.3 Polling procedures

The Front End transmits a poll datagram to the Base Station and the Base Station transmits a poll datagram to the Mobile Station. The datagram structure is given in Figure 14. If the Base Station or the Mobile Station has to transmit data to the Front End, they start after a delay of approximately 15 ms. The modems of the Base Station and the Mobile Station are switched on and they start with the modem synchronisation before the normal SCANET datagram. Approximately 15 ms after the received datagram the Front End and the Base Station send the next poll datagram. Between two poll datagrams there is a delay of approximately 100 ms (see Figure 14).

3.4 Data communication between the Front End and the Base Station

3.4.1 Physical characteristics (layer 1)

Transmission medium:	4 wires
Connection type:	Multipoint
Modem:	
- Modulation:	FSK
- Frequency assignment:	0: 2100 Hz 1: 1300 Hz
- Transmission rate:	1200 bit/s
Communication mode:	Synchronous, half duplex
Hardware interface:	V.24

3.4.2 Data link and the transport network protocol (layer 2–4)

The data link protocol is based on the HDLC protocol.

Primary:	Front end
Secondary:	All base stations
Polling:	Cyclic with individual link address
Maximum frame length:	25 bytes (200 bit)
Maximum frame transmission time:	166.7 ms

3.5 The radio link between the Base Stations and the Mobile Stations

3.5.1 Physical characteristics (layer 1)

Transmission medium:	Radio link
Connection type:	Multipoint
Modem:	
- Modulation:	FSK
- Frequency assignment	1: 1200 Hz 0: 1800 Hz
- Transmission rate	1200 bit/s
Communication mode:	Synchronous, half duplex
Hardware Interface:	V.24

3.5.2 Data link and the transport network protocol (layer 2–4)

The data link protocol is based on the HDLC protocol.

Primary:	Base station
Secondary:	All mobile stations within the range of the base station
Polling:	Cyclic with individual link address. Periodic 'group poll' for logon request
Maximum frame length:	25 bytes (200 bit)
Maximum frame transmission time:	166.7 ms

4 TRC system and Automatic Train Control (ATC)

The Automatic Train Control is a system which automatically switches on the train brakes if the maximum allowed speed is exceeded, the train is moving too fast towards a stop signal ('red light') without proper brake action from the engine driver, or the train is passing a stop signal. The ATC gets all the data from information points along the railway line. The information points are 'passive beacons' (balise) located between the rails (see Figure 15). The balise is totally passive and it is only activated when a train is passing. The ATC radio carrier from the train antenna activates the balise. The balise consists of an antenna and electronic encapsulated in a glass fibre reinforced polyester construction (400 x 500 mm).

The ATC transmission part on the engine consists of a transmitter, receiver, antenna and surveillance. The purpose is to energise the balise and receive the codes. The antenna is located under the train engine and transmits a 27.115 MHz carrier. The carrier is amplitude modulated with a 50 kHz frequency clock signal balise synchronisation. The transmitter power is 10 – 15 W.

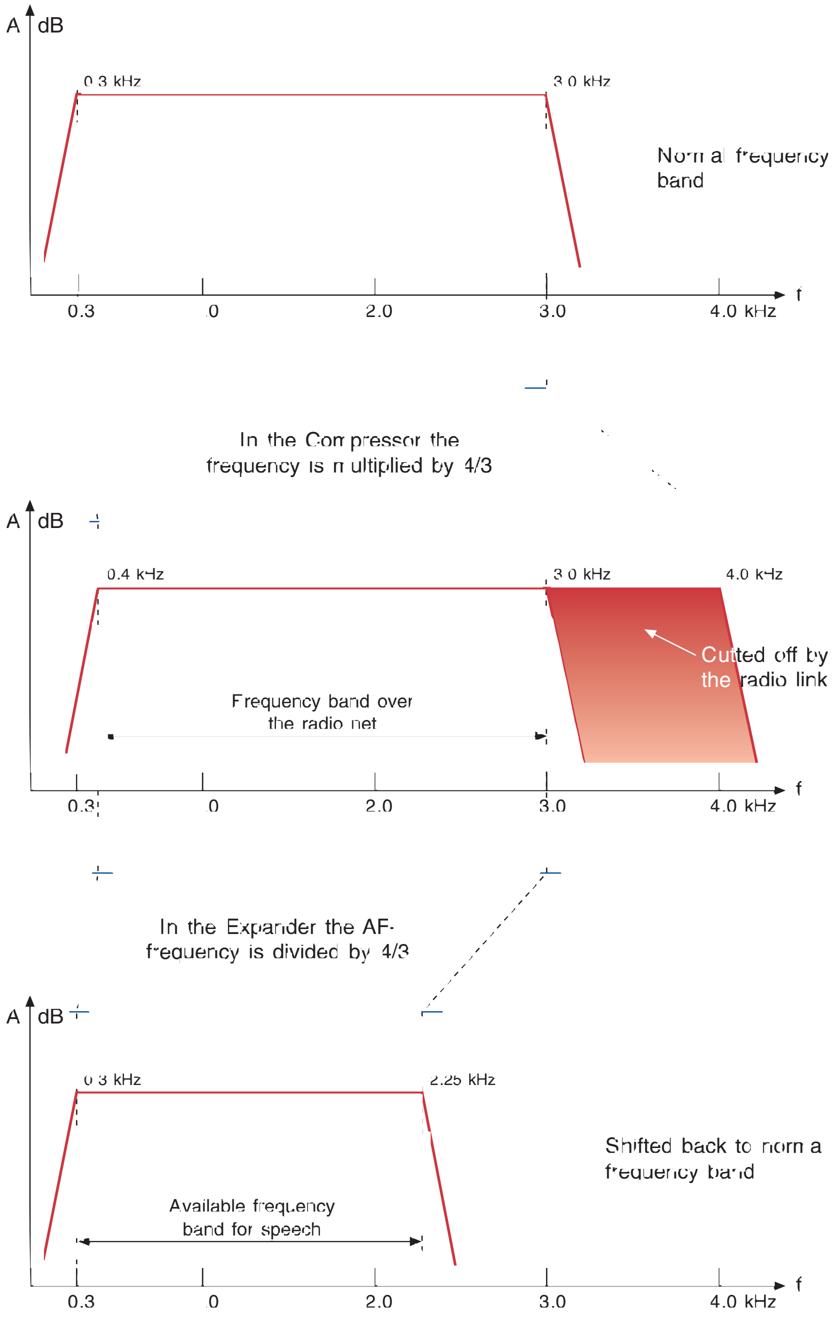


Figure 11 Frequency response after compression/expansion

The signal from the balise to the engine operates on 4.5 MHz with 10 mW radiated power. Each balise is identified with a unique code word (1's and 0's) which is transmitted from the balise when a train is passing. The duration of one bit is 20 µs (50 kHz data rate). Amplitude Shift Keying (ASK) is used for the 4.5 MHz carrier with '1' being carrier switched on and '0' carrier off.

The receiver in the engine receives the 4.5 MHz signal from the balise. It consist of an amplifier, a filter, a detector and an integrator. After each bit period, the integrator voltage is read to decide if the signal is logic '1' or '0'. The surveillance part continuously controls the transmitter level and modulation and it also tests the receiver. The antenna located under the engine consists of an inductive coil for transmitting. It is also a differentially coupled coil for receiving. One coil is for the 4.5 MHz signal and the other is for surveillance.

The train radio system is capable of sending data and the system can also handle data transfer during a conversation. The ATC produces a lot of data that could be

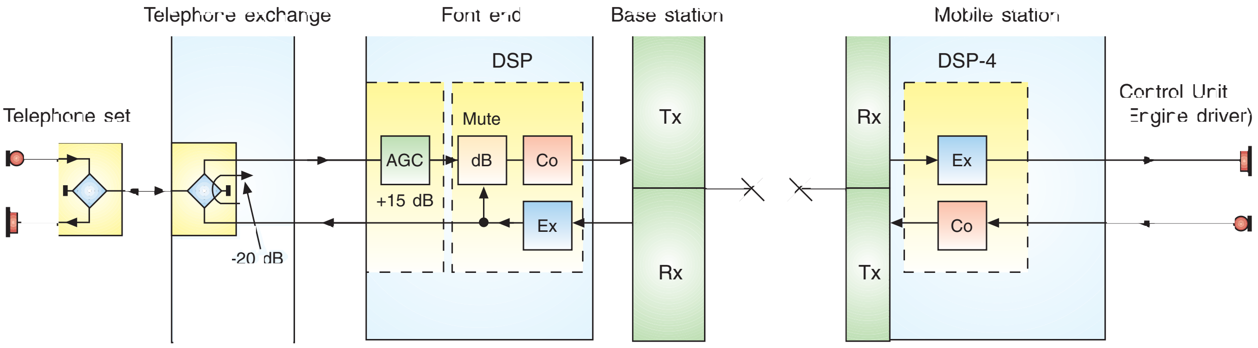


Figure 12 Digital echo cancelling

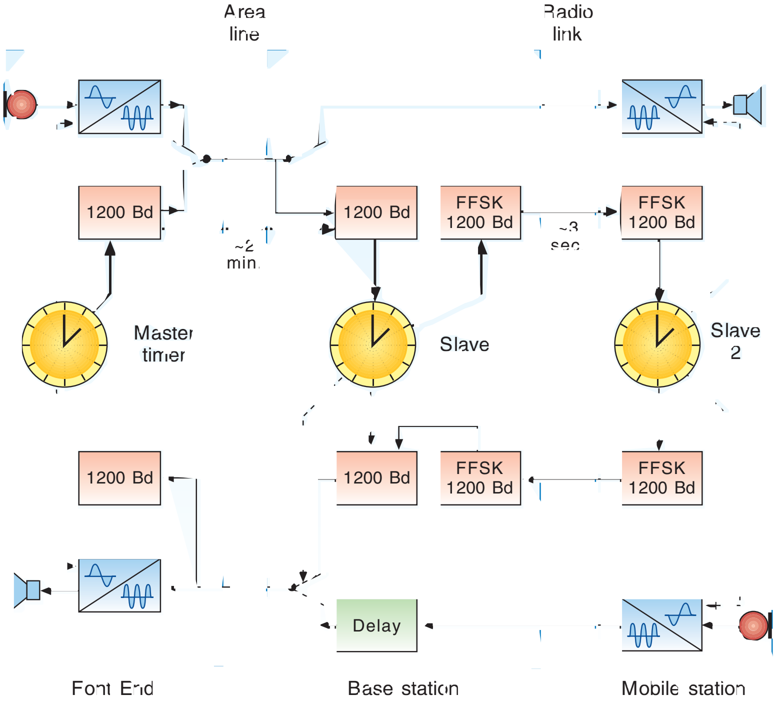


Figure 13 Synchronisation for speech compressor and expander

transferred over the train radio system, but presently only position and area data are transferred from the ATC to the Central Traffic Control.

The train radio system uses two types of balise:

- area change balise
- signal number balise (position balise).

The area change balise controls the change from one radio area to another.

The codes from the balise are sent via the on-board ATC computer to the Mobile Station and the code tells which radio area the train is entering. The Mobile Station gets the information from the ATC system and the radio area is changed automatically. The data are protected with a Hamming (8,4) code. The radio area number is always shown on the command unit display in the driver's cabin.

The signal number balise is located together with the main signal. The railway line is divided into blocks. The typical length of a block is a few kilometres depending on the traffic density and the speed of the train. A main signal (light signal) is always placed at the beginning and at the end of a block. For single track, only one train can be inside a block at the same time. The codes from the balise indicate the number of the next main signal (position of the train). The code used is a modified Hamming (16,11). Bit number 7 and 8 are interchanged to increase the safety. The signal number is sent from the Mobile Station to the Central Traffic Control via the Base Station (see Figure 15). The communication between the ATC and the Mobile Station takes place on a current loop RS232 connection. The interface card in the ATC repeats the latest balise information to the Mobile Station approximately every 4 second, but the Mobile Station does not pass any information to the train radio system before a change occurs. This is done to keep the data transfer in the system at a minimum. At the Dispatch Centre the Hamming code is decoded using a table decoder and presented as a decimal number (signal number) for the train dispatcher.

5 Future GSM train radio system

The UIC (Union Internationale des Chemins de Fer) is undertaking a project called EIRENE (European Integrated Railway Radio Enhanced Network) to define a new standard for a digital radio system for European railway organisations. UIC has decided that the new standard will be based upon the GSM digital cellular standard developed by ETSI. The

ETSI GSM standard will be enhanced to provide a number of features required by the railway organisations. The specification work is being carried out within the GSM phase 2+ standardisation programme. The additional specifications are particularly concerning the voice broadcast, group call and priority set-up services. The radio system is called EIRENE.

The fundamental requirements of the system are that it should

- operate in the 900 MHz frequency band
- be a digital system based on GSM capable of operation with train speeds faster than 500 km/h
- be an open standard enabling competition between manufacturers
- give a broad user base to provide economies of scale.

The special railway requirements are that it should enable

- broadcast calls (for emergency broadcasts over tens of kilometres of the railway line)
- group calls (for local teams and for driver-to-driver group calls over wide areas)
- priority and pre-emption (to ensure operational calls are granted network resources in preference to less important administrative calls)
- fast call set-up (about 1 second for emergency calls)
- GPRS (General Packet Radio Service)
- interpretation of pre-defined short messages
- shunting radio
- multiple driver communication.

A common European frequency allocation is required. The exact up- and downlink frequency band will be decided by CEPT in the near future.

The first test of the system will take place in Germany in 1995. The test is only on standard GSM but with respect to existing services for railway applications. Important test items are the speed requirements and data transfer. Based on this time schedule, the first prototype EIRENE will probably be available in 1997.

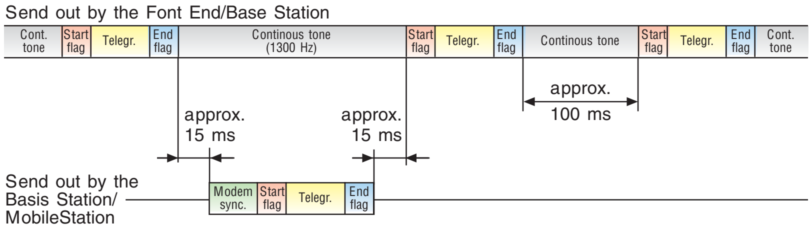


Figure 14 The poll datagrams and the answer of one base station / mobile station

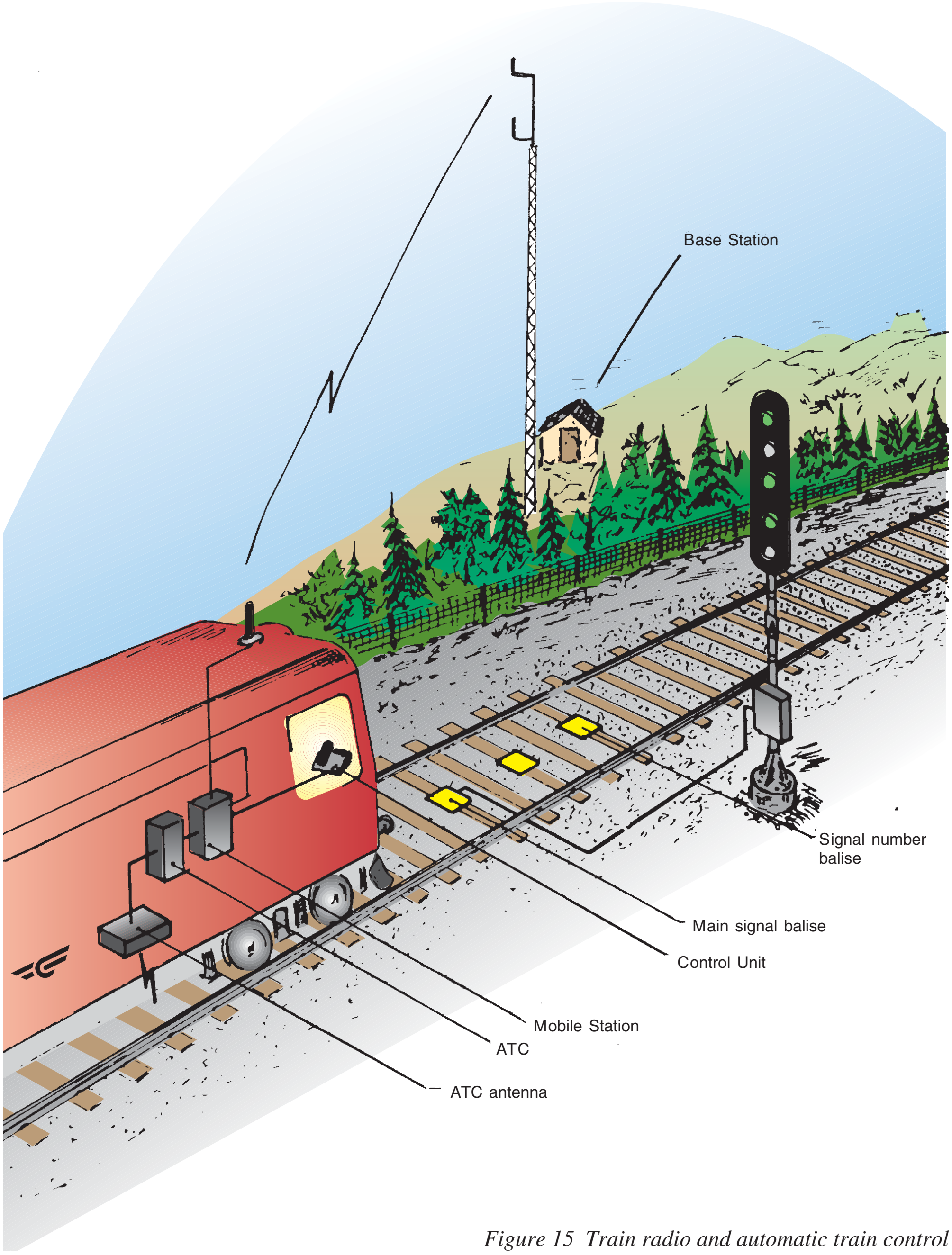


Figure 15 Train radio and automatic train control

Intelligent Networks as a platform for provision of service in GSM and DECT

BY ENDRE SKOLT AND IVAR OLDERVIK

This paper presents personal communications characteristics such as terminal mobility, personal mobility, service mobility and personalised service offering. Furthermore, it shows how existing mobile communications systems may migrate towards next generation systems. In particular, the use of Intelligent Networks (IN) concepts for provisioning of services in the Global System for Mobile communications (GSM) and the Digital European Cordless Telecommunications (DECT) are discussed.

This paper addresses architectural and executional aspects. Management and service creation are not covered.

Introduction

In recent years the mobile communications market has been fast growing, and there is no reason to believe that the growth will stop. Cellular networks such as GSM see a market beyond business customers. In fixed networks, mobility is becoming increasingly important, and users want access to the same services both from fixed and wireless terminals.

What will the future of mobile communications look like?

One major branch of mobile communications will evolve towards a concept named personal communications. A common view is that personal communications will at least contain the following elements:

- terminal mobility
- user/personal mobility
- personalised service offering
- service mobility or interoperability.

The next section briefly presents personal communications characteristics and give examples of the above elements.

Characteristics of personal communications services

Terminal mobility

The first element of personal communications is terminal mobility. Terminal mobility implies that the users should be able to initiate and receive calls using a pocket sized terminal or a handheld terminal wherever they happen to be at the time of the call. Examples of mobile communications systems that support terminal mobility are the cellular networks NMT and GSM, and the wireless access systems DECT and CT2. In the near future there are also plans to offer satellite telephony and data services via handheld terminals.

The GSM system offers the customers full availability and the possibility to make outgoing calls within the area of coverage. The network is structured into cells allocated specific frequencies. The cell radius will typically be a few kilometres.

The wireless access systems DECT and CT2 are characterised by short radio range and high capacity.

Satellite organisations such as Inmarsat and Iridium have invested large amounts of money in order to provide global telephony to handheld terminals. The Iridium contractors have proposed an ambitious plan to launch 66 satellites in low earth orbits to enable global coverage. The focus is mainly on the international business community. Inmarsat P21 claims to meet the demand both from customers in remote areas and the business traveller by launching 10–15 satellites in an intermediate circular orbit. Both organisations say they will have commercial services by the year 2000.

User/personal mobility

Another important element of the personal communications service is user/personal mobility. A telecom subscription should not necessarily be linked to the terminal equipment as in today's public telephone service, where a telephone number is attached to a subscriber line. With the introduction of user mobility, users with personal numbers and personal accounts should be able to share terminals.

Universal Personal Telecommunications (UPT) is the service concept developed in the standardisation bodies to support user mobility in all types of networks (Public Switched Telephone Network (PSTN), Integrated Services Digital Network (ISDN), Broadband ISDN, PLMNs, satellite networks, etc.) by using IN capabilities. UPT phase 1 has been designed for PSTN only. UPT offers a set of customised services, and based on a unique, personal telecommunications number, the user can initiate and receive calls at any public telephone set irrespective of geographical location.

Telenor was the first public network operator in Europe who started UPT trials. Later this year a commercial service will be offered. The only requirement for the service introduced in the Norwegian public network, is that the telephone sets must have DTMF capabilities.

With access to advanced services such as UPT, private numbering plans, screening services, routing based on different criteria, there is a need to allow the user to interrogate and modify service data. IN support network capabilities for this purpose. The user may select items in the service profile based on interaction with announcement machines and DTMF receivers. Users having ISDN-access may be offered more advanced user procedures.

User mobility is supported by GSM. A GSM subscription is linked to the Subscriber Identity Module (SIM), and the mobile phones can be shared by many users. The separation user/terminal implies increased competition among the mobile network operators in such a way that the users can switch between service providers without being obliged to invest in new mobile telephones.

Personalisation of services

It is recognised that implementation of advanced services or supplementary service in PSTN and ISDN lack the flexibility required both viewed from the network operator and the customer. New ISDN services to the customers require new software releases by the vendor, and today's networks are not structured in such a way as to easily offer services tailored to customer specific needs. Since the GSM system has been designed to interwork with PSTN and ISDN, the GSM service providers experience the same inflexibility.

One of the ideas behind IN was to address this problem. Introduction of services should be more effectively handled by deploying software in network entities separated from the switched network. Today, the first IN standard has been finalised, and most of the public network operators in America, Europe and Japan have introduced network architectures based on IN principles. The most common services implemented are advanced Freephone, Universal Access Number, Card Calling services, Virtual Private Networks, Premium Rate, Personal Number, and Televoting.

Service mobility

The last element we will mention as part of personal communications is service mobility. Whether the user accesses the service in the home domain or in a visiting domain, the service should have the same "look and feel" and be available through the same access procedure. By distribution of user data locally to visiting location register (VLR), the GSM system provides service mobility within the area of coverage. In addition to user location information, the VLR contains subscription information and allows the same supplementary service in the visiting and home networks.

In the fixed network, the international telephony service has for many years been available world-wide, and today, end-users in Norway can automatically be connected to end-users in more than 2,000 countries and regions. However, with the demand of more advanced international services the existing interconnections will not meet the need for the future. In that respect, GSM has paved the way for distribution of data in the fixed network. The next IN standard (1996) may include capabilities for distribution of user service logic and user data across public networks, which will allow service mobility. There are many obstacles both technically and regulatory, but we believe that unless data distribution is adopted by the public fixed networks in the near future, they will become mere bit carriers.

Service mobility implies that several networks may be involved in call processing and usage of transmission resources. The user should still only need to relate to one service provider. Single point of contact is a must in service provisioning.

Cordless Terminal Mobility

Introduction

Cordless Terminal Mobility (CTM) may be one of the migration paths towards personal communications. The CTM concept offers both personal and terminal mobility, and can be viewed as a UPT service in a wireless network. There are several reasons why this concept has emerged. Keywords are increased attention to wireless networks and the introduction of personal mobility in the fixed network. The fixed public network operator may see CTM as an opportunity to provide mobility in competition with the conventional mobile network operators. Services traditionally linked to the fixed network may be provided irrespective of fixed or wireless access. On the contrary, the mobile operators may see CTM as a complementary service to GSM. In this case, dual mode terminal will be needed.

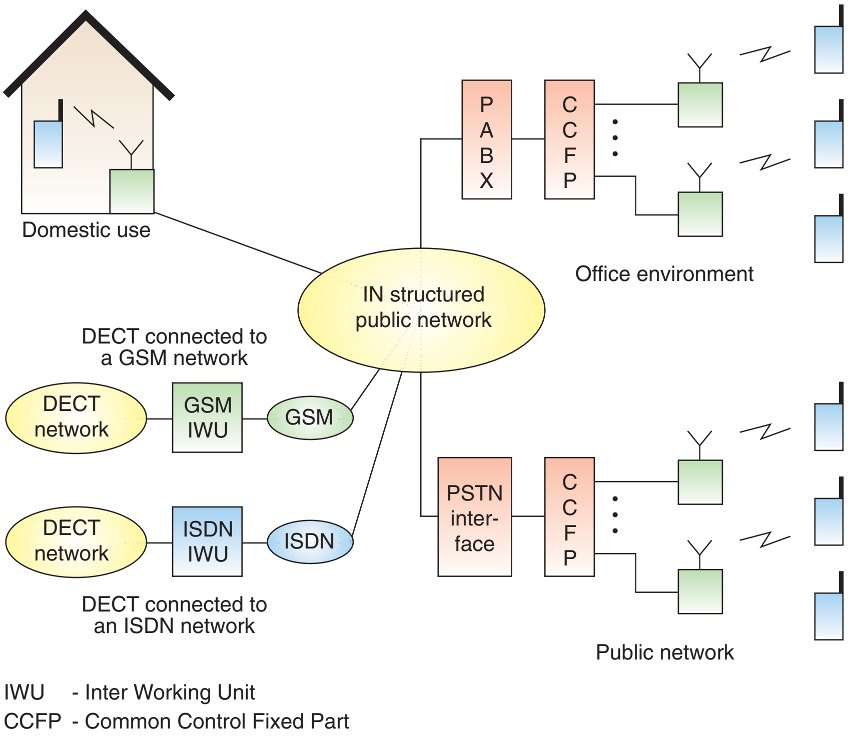


Figure 1 Applications of CTM using DECT wireless access

From the network point of view, work on CTM started in 1994. As for UPT, IN has been chosen as the concept on the network side. However, in some ways CTM does not fit in the work plan and evolution of IN. At an earlier stage, it has been decided that the first and second IN standards should not support terminal mobility. The only way to come around this problem is either to develop a proprietary IN standard in Europe supporting CTM, or design a very limited CTM service.

Applicable scenarios

What type of customer groups may be potential CTM users? With the flexibility of DECT, the following scenarios may become viable:

- Domestic use; user can gain mobility and enhanced services at home
- Office environment; providing mobility, contactability and enhanced services to employees
- Providing coverage and enhanced services in areas like airports, railway stations, city areas, and shopping centres
- Complementary to GSM in areas not supported by GSM. Requires dual-mode terminals.

Roaming between different public domains and public and private domains may also be supported in the near future.

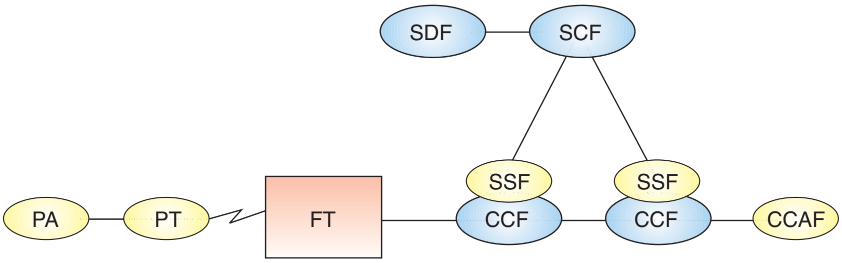
The basic service will include the possibility to make outgoing calls, receive incoming calls and automatic registration of location. In addition, services such as screening of incoming calls, mail box, service profile interrogation and modification, variable routing, follow on, etc., have proved to be interesting customer services.

CTM services and network architecture

In the following we will present some examples of services followed by some architectural considerations.

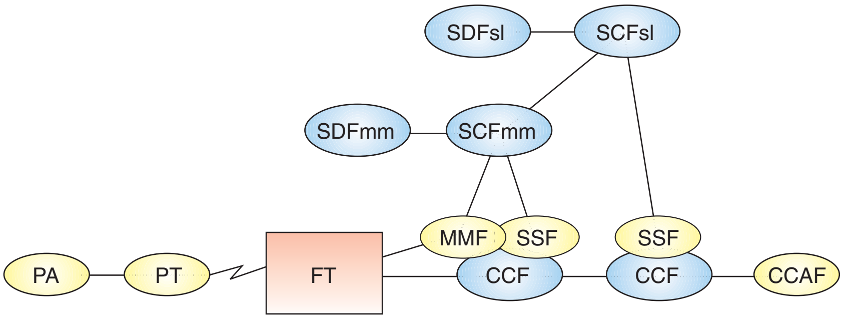
The minimum CTM service set will contain the following procedures:

- Call origination; allows the user to initiate a call from a portable terminal. An authentication procedure should be mandatory in order to validate the terminal.



CCF	- Call Control Function
PA	- Portable Application
SCF	- Service Control Function
SRF	- Specialised Resource Function
CCAF	- Call Control Agent Function
FT	- Fixed Radio Termination
PT	- Portable Radio Termination
SDF	- Service Data Function
SSF	- Service Switching Function

Figure 2 CTM implemented in an IN CS1 functional architecture



CCAF	- Call Control Access Function
CCF	- Call Control Function
PA	- Portable Application
PT	- Portable Terminal
FT	- Fixed Termination
SCFsl	- Service Control Function (service logic)
SCFmm	- Service Control Function (mobile management)
SDFsl	- Service Data Function (service logic)
SDFmm	- Service Data Function (mobility management)
SSF	- Service Switching Function
MMF	- Mobility Management Function

Figure 3 CTM realised in an enhanced IN functional architecture

- Incoming call; allows the user to receive calls based on a personal number. If the user is outside the radio coverage area, or the terminal is switched off, the network may return an appropriate message to the calling user or terminate the call in a mail box.
- Terminal automatic registration; allows the user and terminal to update the network location as they move. If the user is not reachable (switched off, out of coverage), incoming calls could be routed to a backup number corresponding to centralised answering system. When the CTM-user switches on the terminal or enters the coverage area the network will be notified.

Examples of additional services are hunt groups (calls could sequentially be routed to different destinations until the call is answered), variable routing (allowing the CTM-user to set up a profile of routing numbers depending of time of day, day of week, dates), screening of calls based on the calling user's number, or based on a password that has to be inserted by the calling user.

What type of network architecture should be selected in order to support the above requirements? The IN CS1 architecture depicted in Figure 2 may support a simple CTM service. Incoming CTM calls will be triggered in the CCFo/SSFo, and the CTM service logic will be processed in the SCF. User data will be stored in the SDF. Using the CS1 approach, a centralised SCF will both execute customer service logic and CTM mobility management functions such as location registration and terminal authentication. However, if automatic location registration and authentication should be mandatory service features, an enhanced functional architecture would be a better choice (Figure 3).

In Figure 3 the SCFmm, SDFmm, and MMF functional entities have been added to handle mobility management. MMF triggers location registration requests. Terminal location data and authentication will be processed in the SCFmm and stored in the SDFmm.

Wireless access

Both DECT and CT2 are candidate radio access systems for CTM.

DECT is a standard for wireless connection to fixed networks, being adaptable to both existing and future network standards. Networks of interest are PSTN, ISDN, GSM and PABX. DECT, being a replacement for the fixed telephone, should offer the same speech quality as wired networks.

DECT offers a 32 kbit/s ADPCM speech coded channel, and provides high capacity. It is designed to handle 10,000 Erlang/km^2/floor, made possible by microcell structure and available bandwidth from 1880 MHz to 1900 MHz. A base station normally has an indoor range of approx. 50 metres and an outdoor range of approx. 300–400 metres. With directive antennas the range will increase to typically 80–90 metres and several kilometres, respectively. For data transmission DECT supports flexible data rates up to 552 kbit/s range and is capable of handling group III and IV faxes. DECT also provides dynamic channel allocation and seamless handover.

DECT interworking profiles

DECT consists of several profiles describing the services and how they are implemented. Some of these profiles are not yet standardised.

- The Public Access Profile(PAP) supports solutions such as telepoint. With this application DECT opens up for access operators to establish radio access to the fixed network. It is not mandatory in the PAP to support incoming calls.
- The goal of the DECT/ISDN interworking profile is to offer ISDN services in DECT. This profile specifies a range of services like 3.1 kHz telephony, 7 kHz telephony, video telephony, voice band based data transmission (i.e. fax and modems), digital transmission like X.25 over ISDN and other services like group IV faxes, telex, etc.
- The DECT/GSM interworking profile defines an interworking function between a DECT network and a GSM network. This interworking profile offers all roaming scenarios as described in the GSM specification, while handover is only offered inside a DECT network which is limited by the DECT control part. However, handover between DECT neworks is under study in RES3. A DECT network will be connected to the GSM network via the A interface. It is also the intention to offer supplementary services (SS), unstructured supplementary services data (USSD) and short messages services (SMS) on the DECT side.
- The generic access profile (GAP) provides means for mobility functions on the air interface. Any handset can be used for any base station to perform 3.1 kHz teleservice, and both outgoing and incoming calls are supported.
- The data interworking profile supports the use of DECT as a wireless LAN (local area network) with data rates up to 552 kbit/s.

CT-2 is, like DECT, a digital wireless connection to fixed networks. With the implementation of common air interface (CAI), the CT-2 CAI supports inter-operability between base stations and terminals independent of the manufacturer.

The CT-2 is mainly operating in the frequency band 864.1 MHz to 868.1 MHz, but in some countries other frequencies are used. CT-2 supports a traffic capacity of 500 Erlang/km^2, which is well below the capacity for DECT. CT-2 also uses 32 kbit/s ADPCM speech coding, and provides handover (not seamless).

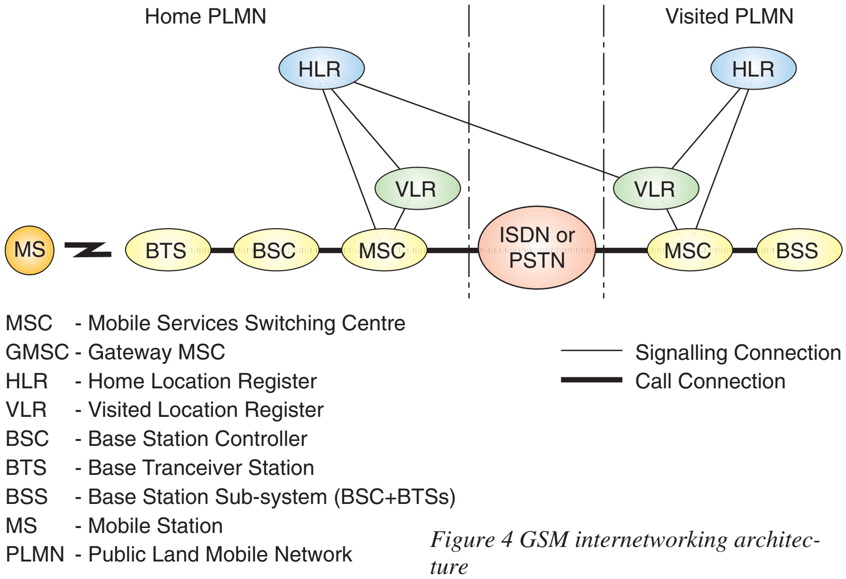

Home PLMN
Visited PLMN
HLR
HLR
VLR
VLR
MS
BTS
BSC
MSC
ISDN or PSTN
MSC
BSS
MSC - Mobile Services Switching Centre
GMSC - Gateway MSC
HLR - Home Location Register
VLR - Visited Location Register
BSC - Base Station Controller
BTS - Base Tranceiver Station
BSS - Base Station Sub-system (BSC+BTSs)
MS - Mobile Station
PLMN - Public Land Mobile Network
Signalling Connection
Call Connection

Figure 4 GSM internetworking architecture

Integration of GSM and IN

Introduction

While the GSM system has shown great success since it was commercially launched 2–3 years ago, it has proved to be rather inflexible with regard to introduction of new competitive services. When the GSM user is roaming, the available services are restricted to a standard set which gives network operators few opportunities in a competitive environment. In addition, aspects such as more efficient usage of transmission resources, fraud control and customised charging need more attention in the evolution of GSM.

Due to the increasing number of subscribers, frequency spectrum will soon not be available to meet the demand for GSM services. Consequently, there will be a need for systems to relieve the GSM radio interface. Candidates are DECT and DCS 1800.

Many GSM operators may have the ambitions to become active players in the household market by implementing GSM in the radio local loop. To be competitive, the fixed network services such as the IN based should be supported by GSM.

GSM characteristics

The reason behind using a cellular structure is to support a large number of subscribers with limited frequency bandwidth. This is accomplished by frequency re-use, which implies that each frequency available is used at several base transceiver stations (BTSs). However, there must be enough distance between the BTSs using the same frequencies to avoid interference.

A mobile user roaming from area to area expects to be reachable. To support communication in a cellular structure implies that there exists means to determine in which cell the mobile user is located. For incoming calls the system has defined location areas. When the mobile station (MS) moves from one location area to another, it informs the network by updating its location. A location area usually consists of several cells, and the MS is paged within its location area when there is an incoming call. The size of the location area is determined as a compromise between how large the areas to page and how much location updating signalling to allow. The network is also able to periodically request location registrations from the MS.

Another important function is to maintain a call when a user moves from one cell to another. This is called handover, and implies that a call has to be switched over to another cell when moving out of the serving cell. This should be done without interrupting the call, and this requires handover signalling procedures. GSM supports the following handover scenarios:

- intra-BSC hand-over; the call is transferred from one BTS to another, but the same BSC is in charge of the call

- inter-BSS/intra MSC; the call is transferred from one BSS to another, but the same MSC is in charge of the call
- inter-MSC; the call is transferred to another MSC/VLR.

The GSM system also offers international roaming. If the home network operator has a roaming agreement with a network operator in another country, the subscriber may use his/her MS for outgoing and incoming calls in the visited PLMN.

The GSM system offers various standardised services. The bearer services implemented in the GSM network supports data rates up to 9.6 kbit/s.

Short message services (SMS) enables alphanumeric messages to be sent to and from a GSM handset. If a handset is not reachable, a message will be stored and sent to the handset when it is reachable.

A number of supplementary services (SSs) have also been specified, such as call forwarding and call barring. New SSs are also standardised in GSM Phase 2. This is a disadvantage of the existing GSM. New services cannot be implemented without an extensive standardisation work. Due to this limitation a new feature called CAMEL[1] is under development. The CAMEL feature may be a first step to integrate IN and GSM, and is a toolbox for operator specific services, with the possibility for tailor made services. However, before describing the CAMEL approach, we will try to understand why IN and GSM integration may be a winning way.

Efficient use of transmission resources

The network procedures needed to support IN service features to a GSM subscriber using the existing infrastructure (GSM network connected to the fixed network via the CCF/SSF – MSC link) will be as follows:

- A GSM subscriber has been recognised by a VLR and the VLR initiates an automatic registration procedure. The HLR returns user data to the VLR. The user data will contain the service profile of the GSM subscriber including a parameter indicating that an IN service feature is subscribed to.
- For mobile terminating calls, the MSC will interrogate the HLR and retrieve the physical routing number (roaming number). When the call detects that an IN service feature is part of the GSM subscription the incoming call will be routed to the CCF/SSF, and a query is sent to the SCF.
- For mobile originating calls the MSC will interrogate the VLR and check the service profile of the GSM user. After successful call processing the MSC will forward the call to the fixed network side. The CCF/SSF will query the SCF, and the IN service feature will be executed.

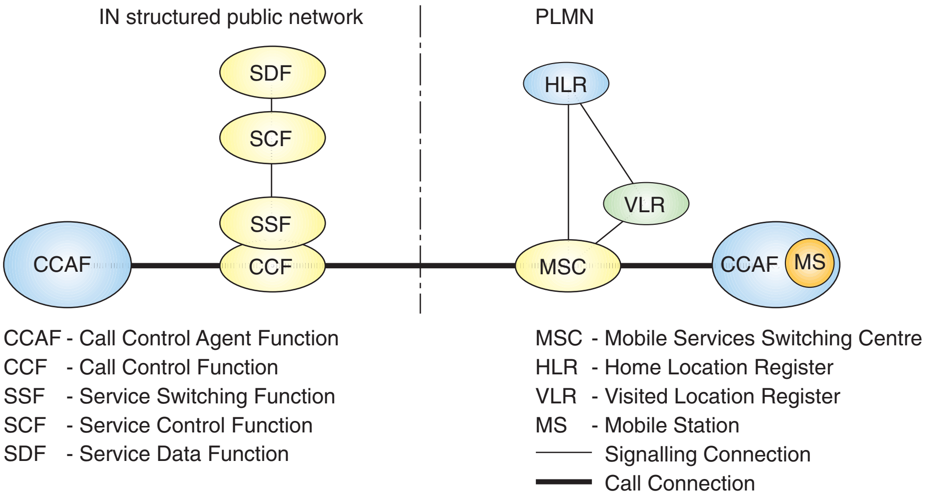

IN structured public network
PLMN
SDF
SCF
SSF
CCF
CCAF
HLR
VLR
MSC
CCAF
MS
CCAF - Call Control Agent Function
CCF - Call Control Function
SSF - Service Switching Function
SCF - Service Control Function
SDF - Service Data Function
MSC - Mobile Services Switching Centre
HLR - Home Location Register
VLR - Visited Location Register
MS - Mobile Station
Signalling Connection
Call Connection

Figure 5 GSM and IN as separate networks

Both of these scenarios illustrate that inefficient usage of network transmission resources may occur. In both cases the calls may be routed via the CCF/SSF in the fixed network, which may cause long and costly transmission paths.

Several ways of combining the GSM architecture and the IN architecture have been proposed. One approach that minimises the impact on existing architectures is to add the IN finite state model to the MSC, which would allow the MSC+ to query the SCF using the IN application protocol (INAP).

A proposal going one step further is to merge the database functions of IN and GSM. A single home database containing the HLR and the SCF/SDF would ease operation and maintenance and make it easier to resolve cases of service interaction between GSM services and IN services.

These above solutions are, however, only relevant if the SCF and the MSC belong to the same network operator.

Charging

Another limitation of the existing GSM is charging inflexibility. Two cases have been mentioned. GSM offers only switch-based charging (charging records are processed and stored in the switch and collected by operation and maintenance centres) which put restrictions on the number of charging records per customer. Too much charging related processing will have impact on the performance of the switch. A second limitation is that roaming subscribers must be charged according to a standard tariff. Customised charging will require additional offline processing.

IN standards give the option to allow generation and control of charging records outside the switching domain, and thereby giving more freedom in charge handling. IN is also developing capabilities for collecting real-time charging information for roaming subscribers.

Fraud

Fraudulent use of the GSM service is an increasing problem. Fraud caused by users roaming into other networks is a specific problem. Operators have already introduced restrictions on subscriber roaming. A more efficient approach would be to monitor calls initiated by roaming subscribers from the home network. Irregular behaviour would be

[1] *CAMEL: Customised Applications for Mobile Enhanced Logic.*

detected very soon, and action could be taken immediately. IN may offer solutions to this problem, however, network integrity and privacy may be another obstacle.

CAMEL

The CAMEL study proposes to implement an IN-like feature to ease the creation of customised services. The objective behind the CAMEL work is to create a “toolbox” which will enable CAMEL service operator specific services. This means that the CAMEL features may be used by network operators or service providers to provide customised services to mobile subscribers also when roaming outside the HPLMN.

The standardisation is currently being executed in ETSI SMG in a phased approach, and stable recommendations on CAMEL phase 1 is expected by third quarter of 1996.

Figure 6 presents a solution for internetworking between home PLMNs and visiting PLMNs. This approach represents a working assumption used in the standardisation body. A final decision on the architecture is likely to take place during September 1995. The term CAMEL INAP is used to indicate that the protocol between gsmSCF and GSSF/VSSF most likely will be based on CS1-INAP with some refinement. This is because CS1-INAP does not support mobility, and hence, some mobility function must be included. No decisions have yet been made on this item.

The gsmSCF contains service logic that will constitute subscriber customised services. The gsmSCF communicates with the underlying switching entities via the CAMEL INAP. When GSM subscribers requiring CAMEL support roams into a visiting PLMN, the gsmSCF will instruct the switching entities in the VPLMN.

Conclusions

In this paper we have described some characteristics associated with future mobile communications such as terminal mobility, personal mobility, service mobility and personalisation of services. Our view is that these characteristics will play a key role in the definition of next generations mobile communications systems.

From the fixed network side, mobility has become an important feature. Personal mobility as provided by the UPT concept will become more wide spread in the market. DECT technology will support some sort of local terminal mobility. The combination of DECT technology and UPT will give a real added value to the users including extended terminal mobility, personal mobility and personalisation of services.

From the mobile network side the main driving forces in the short term are solutions for extended coverage and more efficient usage of the frequency spectrum. However, in the medium term the capability to support customisation of user services will be of vital interest in a competitive market. In this paper we have presented an initiative where IN-principles are introduced into the GSM-environment.

In the past, mobile communication systems have been designed for specific user services such as telephony, paging, and data. In the future we may see more generalised systems. At this moment the telecommunications community is developing a Universal Mobile Telecommunications System (UMTS) which combines features from fixed networks, mobile networks and wireless access. UMTS is scheduled for realisation by 2005. An important objective is that the current developments of CTM and IN/GSM should be harmonised with the UMTS path.

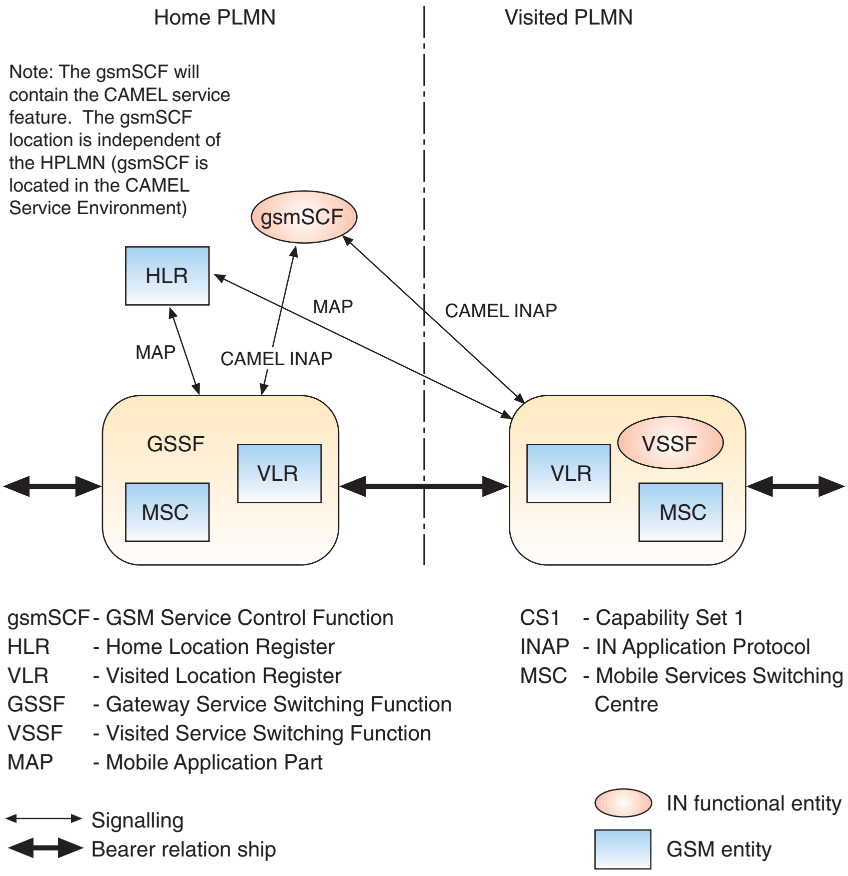

Home PLMN
Visited PLMN
Note: The gsmSCF will contain the CAMEL service feature. The gsmSCF location is independent of the HPLMN (gsmSCF is located in the CAMEL Service Environment)
gsmSCF
HLR
MAP
MAP
CAMEL INAP
CAMEL INAP
GSSF
VLR
MSC
VLR
VSSF
MSC
gsmSCF - GSM Service Control Function
HLR - Home Location Register
VLR - Visited Location Register
GSSF - Gateway Service Switching Function
VSSF - Visited Service Switching Function
MAP - Mobile Application Part
CS1 - Capability Set 1
INAP - IN Application Protocol
MSC - Mobile Services Switching Centre
Signalling
Bearer relation ship
IN functional entity
GSM entity

Figure 6 A possible architecture for the support of CAMEL

HIPERLAN – a decentrally organised wireless LAN

BY CHRISTIAN PLENGE

Wireless local area networks (WLANs) are expected to be a major growing market for the communication network industry in the coming years. Due to their flexibility they will serve as one of the most important communication networks for pico-cellular scenarios. Decentral organisation enables 'ad-hoc' configurations of small WLAN networks and guarantees the adaptability to different scenarios and requirements. HIPERLAN (HIgh Performance Radio Local Area Network) is the European standard [5] for wireless LANs, which is currently getting the official status as "draft standard". HIPERLAN will provide a data rate of about 20 Mbit/sec and different frequency bands to be used in adjacent systems. The standard comprises the functionality for the decentral organisation of the medium access and of the network management for forwarding, security and the use of priorities. This paper will introduce the basic structure of the HIPERLAN standard as it is a standard for the medium access layer. Main focus will be on the decentral channel access mechanism (CAM) which is the crucial part for the performance of the system. The most important mechanism described is under discussion for the final version. Its performance will be evaluated using stochastic simulations. Problems that are typical for decentrally organised networks like hidden nodes and capture, will be analysed due to their influence on the system performance.

1 Introduction

Growing needs for the mobility of people and their equipment like laptops or PDAs in office environments, in conferences and meetings, as well as a high demand for flexibility of installed business communication networks are the initial keys to the market for wireless local area networks. In the future they will extend existing cabled solution if mobility shall be provided, substitute them where flexible adaptable topology is required and will be used widely for new implementations. But such systems will only be commonly used world-wide if a standardised system is available. HIPERLAN (HIgh PERformance Local Area Network) is the European alternative to the American Standard for WLAN (IEEE 802.11). Without special licence HIPERLAN can be used as broadband, flexible 'ad-hoc' LAN and combined with other LANs. Without a central base station HIPERLAN enables direct computer-computer communication. The advantage of this network structure compared to existing 'wireless LAN' solutions with central access points is therefore HIPERLAN's flexibility and mobility. At conferences and in project meetings the decentral organisation of HIPERLAN connects computers without infrastructure facilities. Each laptop carries its own part of the HIPERLAN network. In this application of HIPERLAN the importance of a European standard becomes obvious. Customers will only buy their part of the network if they are sure that everybody will use the same system. On the other hand, re-design of wired information systems in office environments becomes uncomplicated with HIPERLAN because components can be added, moved and removed easily – without a breakdown of the network. If one node in a HIPERLAN has got an access to existing wired LANs and the Internet, it acts as a relay. For the user the HIPERLAN appears as part of the wired network. Such scenarios and other possible HIPERLAN topologies are depicted in Chapter 2.

HIPERLAN supports a broad variety of applications, which may use data, voice and video. Time bounded services are realised by dynamic priority levels for different data types within the channel access. Uni-, multi- and broadcast transmission can be used while HIPERLAN manages relaying, forwarding and acknowledgement decentrally. The HIPERLAN standard includes access, routing and forwarding functions for that case which will be introduced in Chapter 4. In addition, these functions provide the extension of a HIPERLAN node's communication range beyond its own communication range. Every node within one HIPERLAN can be addressed, if at least one possible routing path within the HIPERLAN exists.

Table 1 Carrier frequencies

Carrier number	Centre frequency
0	5,176.4680 MHz
1	5,199.9974 MHz
2	5,223.5268 MHz
3	5,247.0562 MHz
4	5,270.5856 MHz

HIPERLAN uses the frequency band between 5.15 and 5.30 GHz (and intends to use also 17.1 – 17.3 GHz). To avoid disturbance and collisions between adjacent HIPERLANs this band is divided into 5 frequency channels (see Table 1), each with a data capacity of 23.53 Mbit/s.

The channel access is organised with decentral time multiplexing between all HIPERLAN nodes, using special protocols based on Carrier Sense Multiple Access with Collision Avoidance (CSMA/CA). Different solutions have been under discussion within the standardisation process. In Chapter 5 the basic approach and the standardised protocols will be described in detail. The performance evaluation of this decentrally organised medium access comprises the basic network configuration of one HIPERLAN with some hidden nodes as asynchronised traffic sources (Chapter 6). The interesting performance parameters are the overall data capacity under these conditions as well as the suitability for time-critical services. For this the delay of data transmission was analysed in detail.

2 HIPERLAN network topologies

Every HIPERLAN network is divided into virtual subnetworks, so called cells, with a global unique HIPERLAN identifier (HID). Every entity (node) is characterised by a global unique node identifier (NID).

If one station sends a frame, the radio channel is busy for all this frame sensing stations within this cell. The stations outside the cell are not able to receive the frame and regard the channel as free. Stations which are not members of the cell, but within its radio range, are interfered by the cell. Also from these hidden stations the communication from the participants in the cell is interfered. Ideally, the cells are round. But in reality the cells get an arbitrary form through reflections, for example from walls. So the planning of wireless local networks is difficult, especially the global development of an area. A cell consists of at least one transmitter and one receiver and will be defined as the room in which all members use the same frequency. All stations in a cell have equal access and can communicate with each other. The stations in a cell have always radio contact and so they can be mobile in the cell. Only the size of the cell is a limitation for the user. The size of the cell is limited by the transmitter power and by the quality of receiving from the transmitter and the receiver. HIPERLAN will be used for in-house networking with a maximum range of 50 m. If a station leaves the radio range from its HIPERLAN, this station will be inaccessible for other members of the cell. In the following some different structures of cells are shown, which are an area of application for HIPERLAN.

2.1 Ad-hoc networks

This is a quite new concept which relates to the establishment of a network on an ad-hoc basis, which means only when needed. Ad-hoc networks can be formed by a number of stations and will be established for instance in a meeting where participants take notes directly on their laptop, or will bring their presentation material to the meeting directly in electronic form. These networks make use of a locally available HIPERLAN equipment, such as printers or large screen displays, to show or distribute their material to all participants of the meeting. This network should be able to be set up fast, and would perhaps only last for the duration of the meeting. For this purpose, a HIPERLAN can be established without any support from a backbone or any other wired infrastructure. The main advantages from these ad-hoc networks are the flexibility and especially the mobility resulting in many possible applications.

2.2 Independent cells

Two cells, A and B (Figure 2), which are not connected with each other and nobody from cell A is in radio range from a station of cell B, are called independent cells. Also when the stations from the two cells use the same frequency for communication, it is presumed that the cells do not use the same medium for communication, and so nobody interferes a member from the other cell.

2.3 Overlap cells

If some stations of a HIPERLAN A are within radio communication range of some members of B (Figure 3), the members in this area use the same medium for communication and the same transmitting capacity.

In an overlap situation two effects can occur:

- The transmitter in the different cells use the same frequency band, so the frequency band cannot be used optimally, because the stations interfere with each other's stations (hidden stations).
- A station receives data packets from several cells with different HID. Each received data packet will be evaluated, but only those with their own HID will be accepted. In this case the throughput of a cell falls.

By an introduction from several frequency channels this effect can be reduced. HIPERLAN supports up to five channels from 5.15 to 5.30 GHz (and perhaps 17.1 to 17.3 GHz).

2.4 Dipole cells

Dipole cells can be used to connect two buildings. This can be realised when a station is networked with a station of the other cell, but the stations are also connected with a directional antenna. For this the stations of each cell must use the same frequency to transmit and receive data packets.

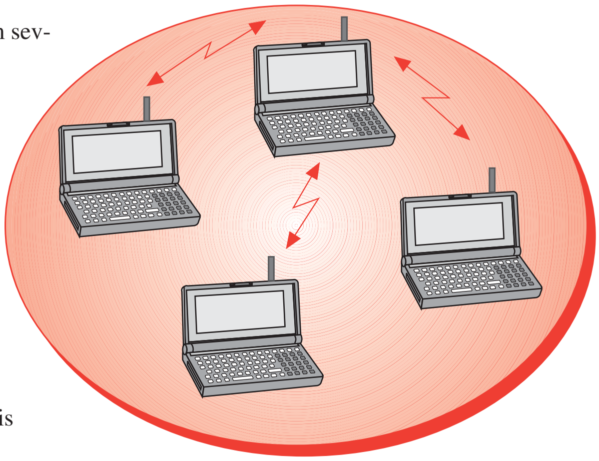


Figure 1 Flexible cell structure

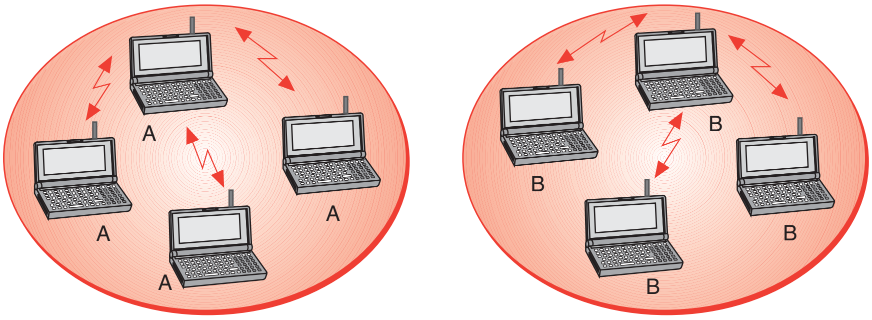


Figure 2 Independent cells

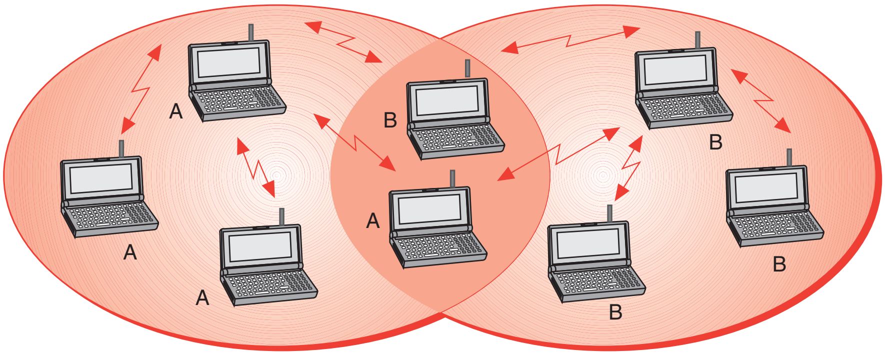


Figure 3 Overlapping cells

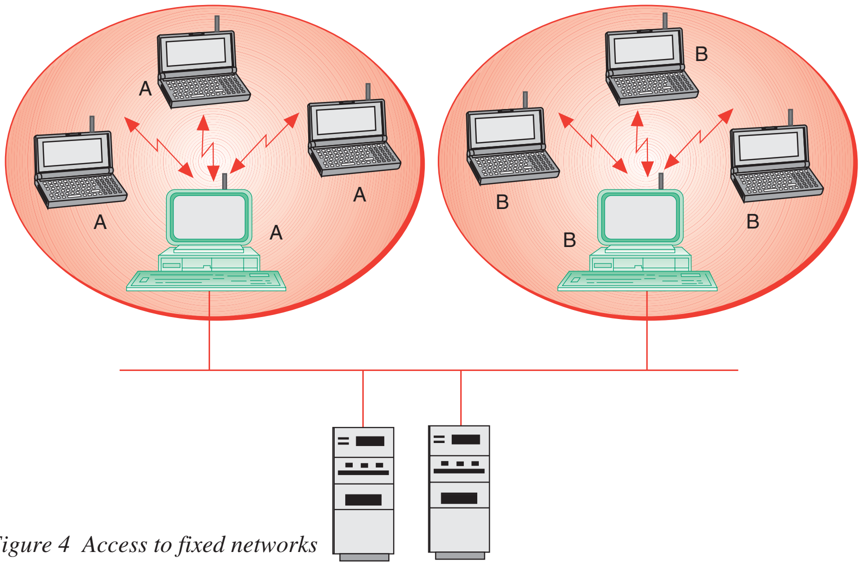


Figure 4 Access to fixed networks

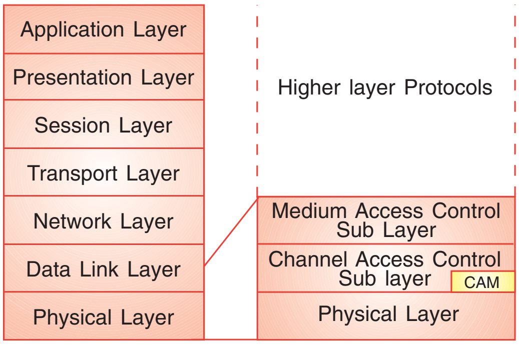


Figure 5 OSI and HIPERLAN reference model

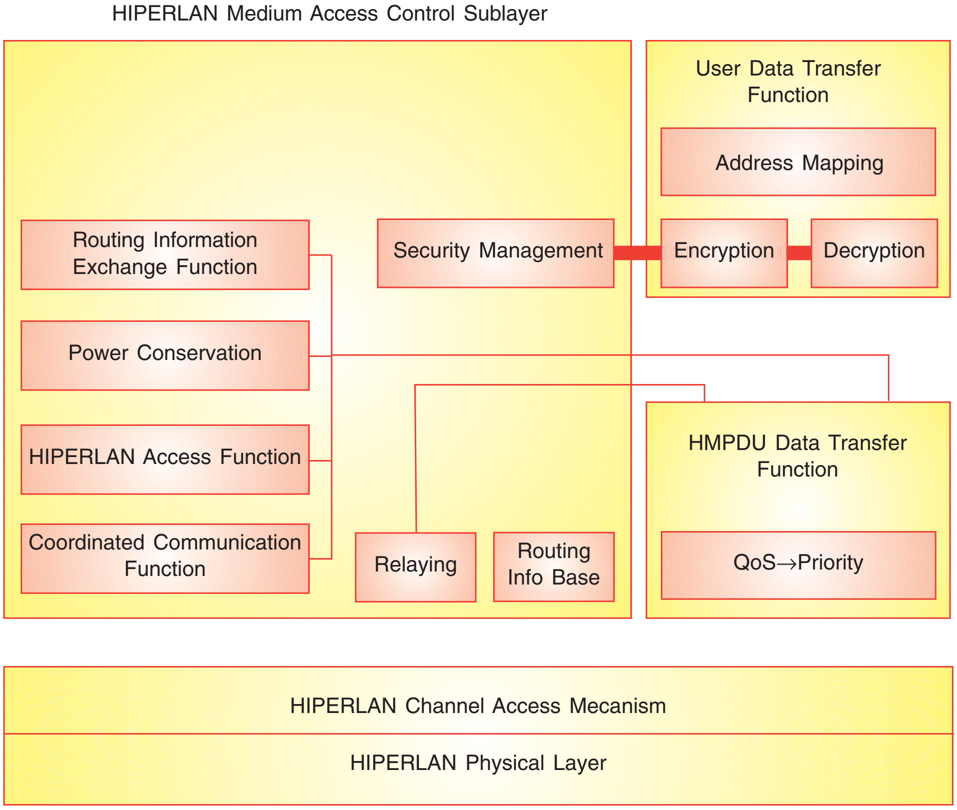

HIPERLAN Medium Access Control Sublayer
User Data Transfer Function
Address Mapping
Routing Information Exchange Function
Security Management
Encryption
Decryption
Power Conservation
HIPERLAN Access Function
HMPDU Data Transfer Function
Coordinated Communication Function
Relaying
Routing Info Base
QoS→Priority
HIPERLAN Channel Access Mecanism
HIPERLAN Physical Layer

Figure 6 System coordination function block diagram

2.5 Connection to a wired LAN

Through a coupling with a wired LAN the stations can use fax, email or an access to a central server. The stations which have an access to a wired LAN must forward the data packets, so they have the function of a router (Figure 4).

3 HIPERLAN reference model

The HIPERLAN reference model is based on the ISO/OSI reference model. The medium access control sublayer from HIPERLAN is subdivided. The part from the MAC-sublayer for the medium access will be separated from the part for the decentral organisation. In the HIPERLAN reference model the part for the organisation is in the MAC-sublayer and the access on the radio channel in the channel access control sublayer (CAC). In this layering model, HIPERLAN applications are outlined as higher layer protocols above the HIPERLAN layers. The mapping of the HIPERLAN reference model to the OSI reference model is shown in Figure 5.

4 HIPERLAN Medium Access Control – MAC

The functional building block of the MAC layer [3] (Figure 6) can be distinguished between internal MAC functions that represent the decentral organisation of the HIPERLAN and the data transfer functions. While the User Data Transfer Function is responsible for the management of data from the HIPERLAN user, the HMPDU Transfer Function multiplexes internal MAC as well as external data packages into one data stream for the CAM. The following paragraphs will briefly introduce the functionality of some important MAC coordination functions.

4.1 HIPERLAN identification scheme

In the HIPERLAN identification scheme, each HIPERLAN shall be assigned a numerical HIPERLAN identifier and a character-based HIPERLAN name. The HIPERLAN identifier is used by the

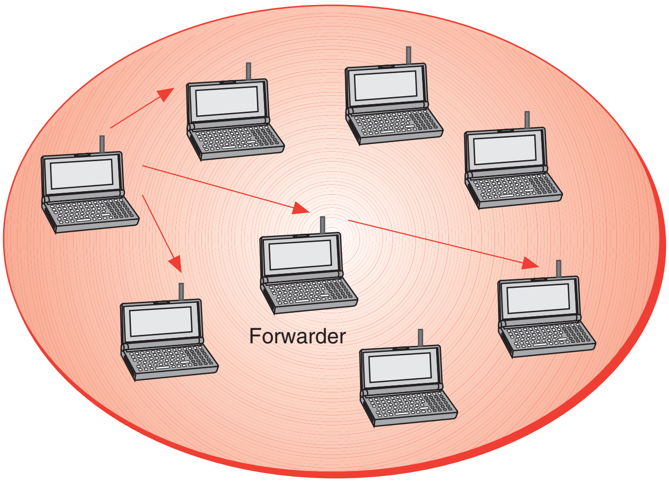


Figure 7 Broadcast relaying

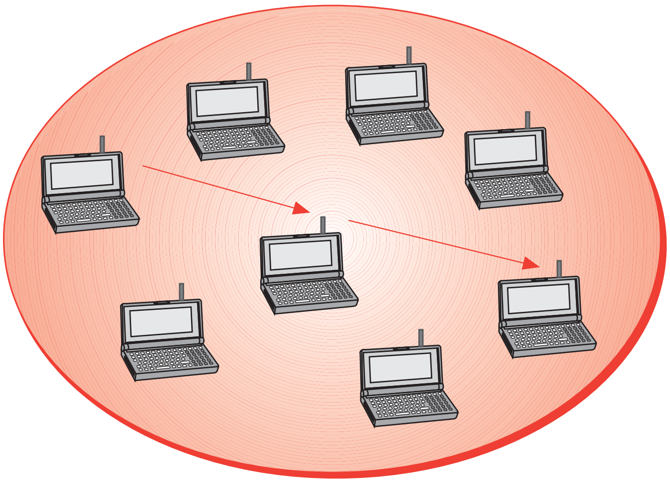


Figure 8 Unicast relaying

HIPERLAN MAC protocol as the means to distinguish between the HIPERLAN MAC communication belonging to different HIPERLANs. If the HIPERLAN identifier is assigned dynamically to take the potential HIPERLAN overlap situation into account, mingled HIPERLAN MAC communication from indistinguishable HIPERLANs is less likely to occur. The HIPERLAN MAC protocol reserves the use of a special HIPERLAN identifier for communicating with the neighbouring HM-entities belonging to any HIPERLAN. The HIPERLAN name may identify a HIPERLAN meaningfully and is used solely by HMS-users as the means to explore the available HIPERLANs. It would be valuable and practical for the HIPERLAN name to be assigned according to the purpose of HIPERLAN deployment.

4.2 HIPERLAN access function

The HIPERLAN access function is defined to explore the HIPERLAN communication environment. The various logical communication activities supported by the access function are:

- *find.* A specific HIPERLAN known by name is found by retrieving the HIPERLAN identifier associated with that name.
- *create.* A HIPERLAN identifier which is not in current use in the communication environment should be chosen as the HIPERLAN identifier associated with the name of the HIPERLAN to be created.
- *destroy.* A HIPERLAN is implicitly destroyed when no HM-entities use the HIPERLAN identifier.
- *join.* An HM-entity implicitly joins a HIPERLAN by retrieving the HIPERLAN identifier associated with the HIPERLAN name and the keys used in encryption and decryption, if any, and transmitting or receiving HMPDUs using that HIPERLAN identifier.
- *leave.* An HM-entity implicitly leave a HIPERLAN by ceasing to use the HIPERLAN identifier in any HMPDU it transmits or receives.

Table 2 HMQoS and channel access priority mapping

NRML	CAP	
	High user priority	Normal user priority
NRML < 10 ms	0	1
10 ms ≤ NRML < 20 ms	1	2
20 ms ≤ NRML < 40 ms	2	3
40 ms ≤ NRML < 80 ms	3	4
NRL ≥ 80 ms	4	4

4.3 Routing information exchange function

Multihop relaying is used by the HIPERLAN MAC protocol to extend the HIPERLAN MAC communication beyond the radio range. MSDUs may be relayed in two different modes: broadcast relaying and unicast relaying. Broadcast relaying (Figure 7) distributes an MSDU to all the HM-entities in a HIPERLAN. It is employed to support multicast or broadcast MSDU transfers. Although it is expensive in terms of data capacity, broadcast relaying may also be used to support unicast MSDU transfers in the absence of sufficient routing information. Unicast MSDU transfers (Figure 8) are more efficiently supported by unicast relaying. Unicast relaying relays an MSDU towards its destination along a route through successive hops. Therefore, unicast relaying is only possible when sufficient routing information is available. This "Routing Information Base" is established decentrally by exchanging special messages between all HM-entities. Each HM-entity shall be either a forwarder or a non-forwarder. A non-forwarder never forwards an MSDU that it has received.

4.4 HMPDU transfer function

At the beginning of every transmission cycle the data packet with the highest channel access priority is determined from the HMPDU transfer function and forwarded to the CAC sublayer. With the residual lifetime, it indicates the amount of time remaining with respect to an MSDU's specified MSDU lifetime and the total elapsed lifetime, divided with the number of hops to reach the receiver,

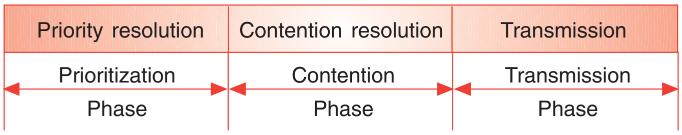


Figure 9 Channel access cycle

the normalised residual MSDU lifetime (NRML) is calculated.

NRML =
residual lifetime / number of hops (1)

Then the channel access priority can be determined by Table 2. Data packets with a high user priority have a higher channel access priority than data packets with a low user priority. Additionally, the highest channel access priority is only reserved for data packets with high user priority. So the time bounded services are supported. The channel access priority is made topical in every transmitting cycle.

5 HIPERLAN Channel Access Control – CAC

The HIPERLAN MAC protocol relies on the provision of a hierarchically independent CAM by the HIPERLAN CAC sublayer to support time-bounded communication. Therefore, the CAM specified by a HIPERLAN CAC protocol shall adhere to the principle of non-pre-emptive priority multiple access (NPMA), which provides hierarchical independence of performance by means of channel access priority such that the performance of data transmission attempts with a given channel access priority is not degraded by those with lower channel access priority. NPMA operates in channel access cycles in the channel. Channel access cycles begin following the end of the previous channel access cycle or at any time after, and for the duration which the medium is considered free. In NPMA, each data transmission attempt is associated with a channel access priority and is synchronised in alignment with the channel access cycles in the channel. As illustrated in Figure 9, an NPMA channel access cycle has three phases: the prioritisation phase, the contention phase and the transmission phase.

A channel access cycle starts with the prioritisation phase, during which priority resolution is performed according to a priority resolution scheme. The priority resolution scheme ensures that only those data transmission attempts with the relatively highest channel access priority will survive the prioritisation phase. In addition, the priority resolution scheme is to be non-pre-emptive, so that only data transmission attempts ready at the start of a channel access cycle may contend for channel access in that channel access cycle, and no new data transmission attempt, even with a potentially higher channel access priority, is allowed to interfere during that channel access cycle.

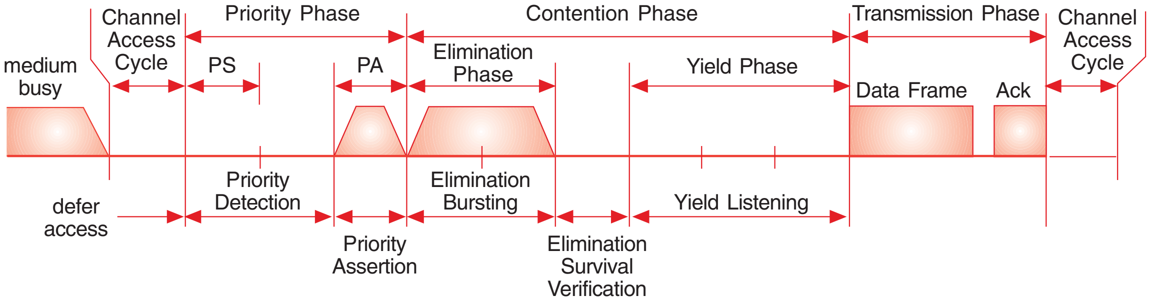


Figure 10 EY-NPMA

The prioritisation phase is immediately followed by the contention phase, during which only the data transmission attempts surviving the prioritisation phase contend for the right of transmission according to a contention resolution scheme. The contention resolution scheme ensures that each surviving data transmission attempt has a statistically equal chance to gain the right of transmission. Finally, in the transmission phase, all the data transmission attempts that have obtained the right of transmission in the contention phase transmit their data.

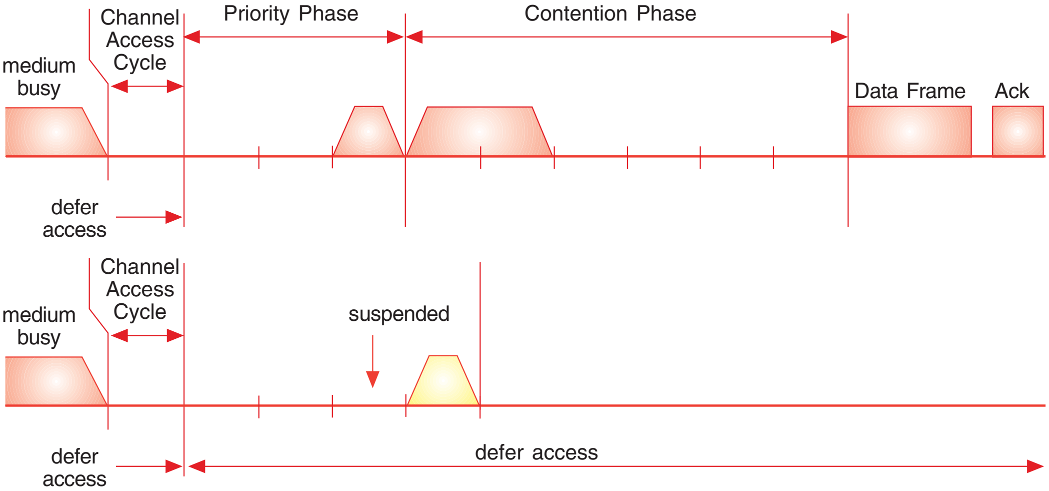


Figure 11 EY-NPMA priority phase

The EY-NPMA is based on the principle of non-pre-emptive priority multiple access (NPMA), which operates with a prioritisation phase, a contention phase and a transmission phase. It uses a combination of the elimination scheme and the yield scheme for contention resolution. The objective of the elimination scheme is to eliminate as many as possible, but not all, contending nodes from competing for the right of transmission. It provides a low, quasi-constant channel collision rate independent of the number of contending nodes. With appropriately operating parameter settings, it has a statistical property that the number of surviving nodes after the elimination scheme is dominantly 1, to a much lesser extent 2 or 3, and very unlikely to be more than 3 [2]. The yield scheme's objective is to complement the elimination scheme. It is designed particularly to further resolve contention between any remaining contending nodes surviving from the elimination scheme. The EY-NPMA activities in a channel are illustrated in Figure 10.

The priority resolution scheme provides hierarchical independence of performance between loads at different channel access priority levels. The operation of the priority scheme is outlined below:

There are a total of m_{CAP} channel access priority levels, which are numbered from 0 to (m_{CAP-1}), with 0 denoting the highest channel access priority level. Prioritisation slots are used for prioritisation between different channel access priority levels. The duration of the prioritisation slot interval is denoted by i_{PS}. Priority resolution takes place by means of priority detection and priority assertion. A contending node, whose frame has a channel access priority *n*, shall listen for *n* prioritisation slot intervals. If the channel is sensed idle during the *n* prioritisation slot intervals, the node asserts the channel access priority by transmitting immediately a burst for the duration of the priority assertion interval, i_{PA}. Otherwise, the node stops its channel access attempt in this channel access cycle. If the prioritisation phase ends with a priority assertion for channel access priority *m*, the duration of the prioritisation interval is *m* prioritisation slot intervals. In principle, at least one contending node will survive the prioritisation phase.

Elimination slots are used for elimination bursting. The duration of the elimination slot interval is denoted by i_{ES} and the duration of the channel access cycle synchronisation interval is denoted by i_{CS}. The duration of an individual elimination bursting is bounded inclusively between 0 and m_{ES} elimination slot intervals, with the probability of bursting in an elimination slot interval being p_E. Accordingly, the probability of an individual elimination bursting to be *n* elimination slot intervals long, $P_E(n)$, is given by:

$$p_E(n) = \begin{cases} (p_E)^n \cdot (1-p_E) & 0 \le n < m_{ES} \\ (p_E)^{m_{ES}} & n = m_{ES} \end{cases} \tag{2}$$

The elimination scheme resolves contention by means of elimination bursting and elimination survival verification. A contending node transmits a burst to eliminate other contending nodes, and then listens to the channel for the duration of the elimination survival verification interval, i_{ESV}, to verify if it is eliminated by other contending nodes. A contending node survives the elimination phase if and only if it senses the channel idle during its elimination survival verification interval; otherwise, the node is eliminated and withdraws from the competition for the right of transmission in the current channel access cycle. When

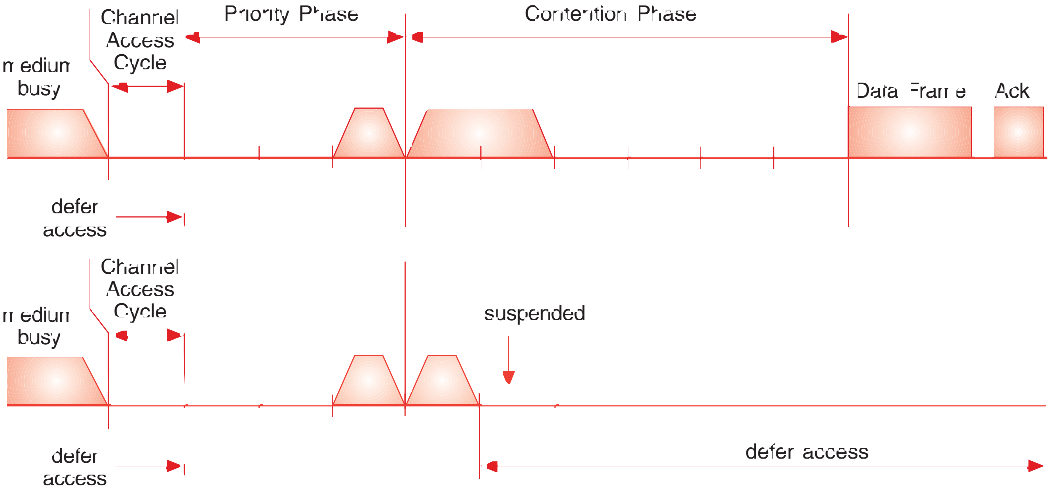


Figure 12 EY-NPMA elimination phase

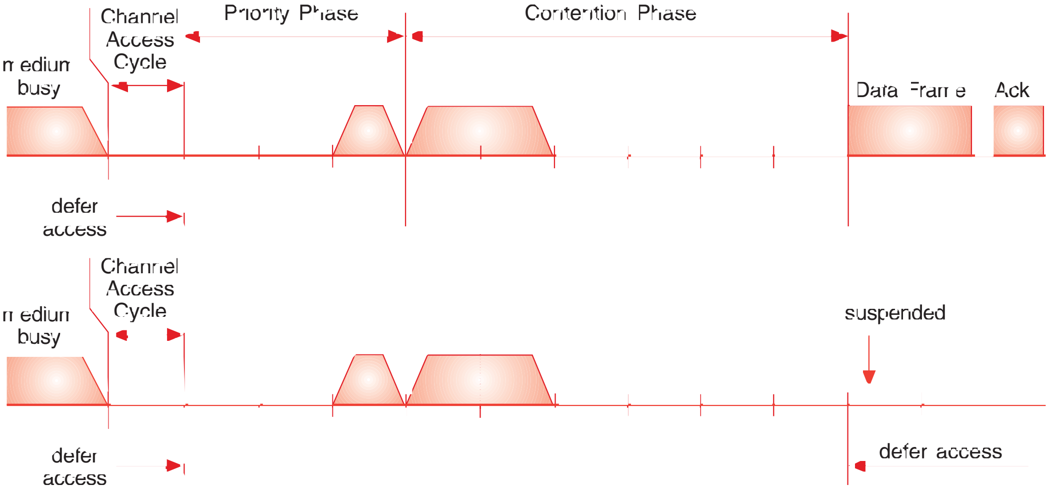


Figure 13 EY-NPMA yield phase

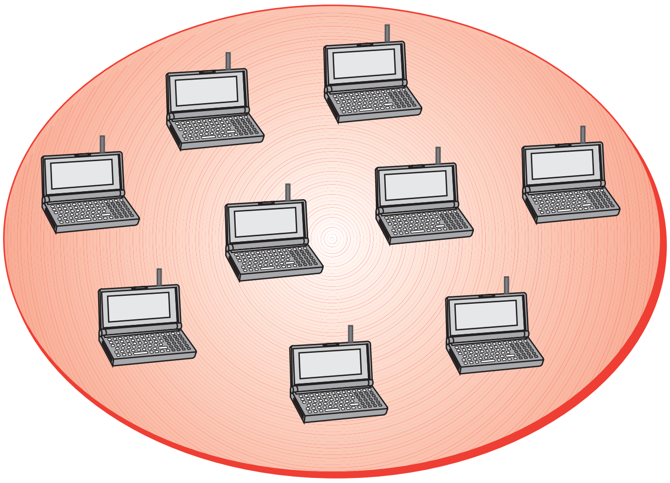


Figure 14 Ad-hoc net

the elimination phase ends, the duration of the elimination interval is determined by the longest elimination bursting among all contending nodes. In principle, at least one contending node will survive the elimination phase.

The yield scheme operates as a supplement to the elimination scheme in the contention resolution phase. Yield slots are used for yield listening. The duration of the yield slot interval is denoted by i_{YS}. The duration of an individual yield listening is bounded inclusively between 0 and m_{YS} yield slot intervals, with the probability for an individual yield listening to be n yield slot intervals long, $P_Y(n)$, is given by:

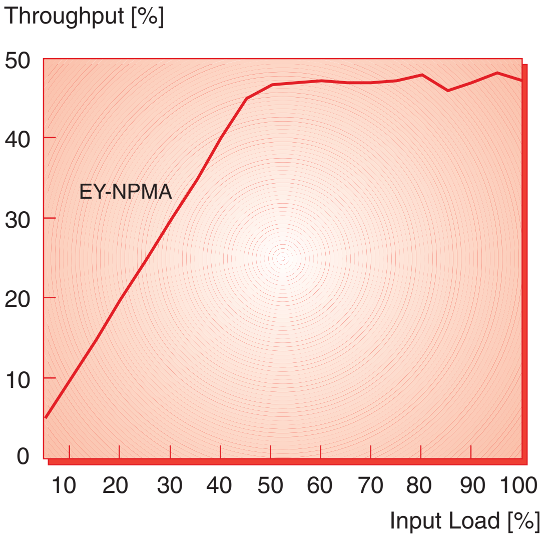


Figure 15 Throughput

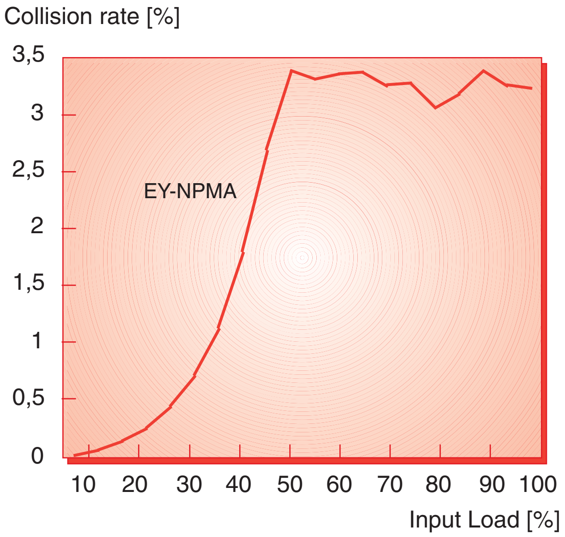


Figure 16 Collision rate

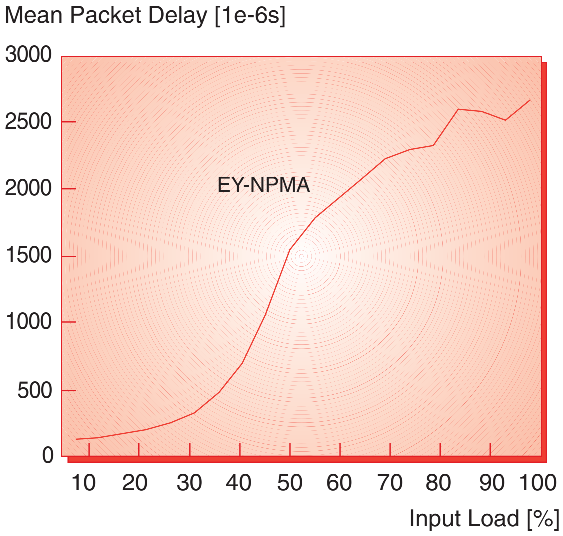


Figure 17 Mean packet delay

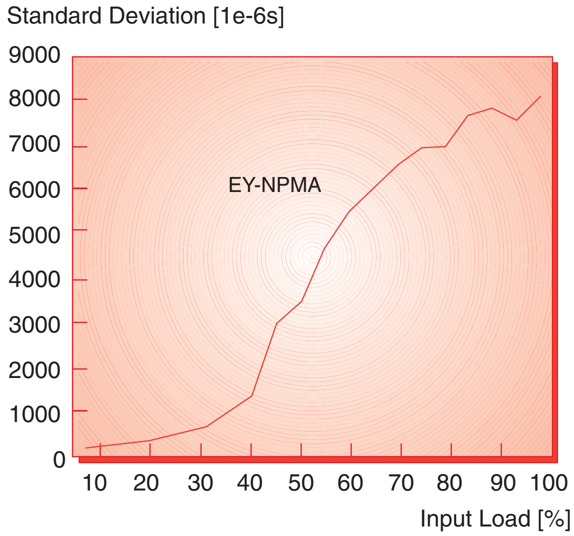


Figure 18 Standard deviation of mean packet delay

$$p_Y(n) = \begin{cases} (p_Y)^n \cdot (1 - p_Y) & 0 \leq n < m_{YS} \\ (p_Y)^{m_{YS}} & n = m_{YS} \end{cases} \tag{3}$$

The yield scheme resolves contention by means of yield listening. A contending node survives the yield phase if and only if it senses the channel idle during its yield listening; otherwise, the node yields to the other contending nodes and withdraws from the competition for the right of transmission in the current channel access cycle. When the yield phase ends, the duration of the yield interval is determined by the shortest yield listening among all contending nodes. In principle, at least one contending node will survive the yield phase.

The CAM provides error control through positive acknowledgement and re-transmission of unacknowledged frames. Acknowledgement is only used for point-to-point transmission and not for broadcast and multicast transmissions. If the transmitter of a frame does not receive an acknowledgement in an ack-time the frame is collided. In this case the frame must re-transmit.

6 Evaluation of the HIPERLAN performance

This chapter shall give an impression of the performance of HIPERLAN networks. The results presented were derived by stochastic simulations applying a realistic HIPERLAN model within the system simulator SIMCO3++/HIPERLAN of the department for Communication Networks at the University of Aachen [1] [4]. The goal was the analysis of the decentral organisation in different scenarios, especially as one cell and with overlapping cells with hidden stations. These basic simulation scenarios are depicted in Figures 14 and 19.

To investigate the channel access mechanism a cell (20 m * 20 m) with eight stations (Figure 14) was simulated. The stations generated data packets with a length of 800 bytes (267 µs). The acknowledgements have a length of 200 bytes (68 µs) and were expected within an (ack time) of 150 µs. Simulations with an input load of 5 % to 100 % (23,5294 Mbit/s) were enforced. All simulations have a duration in which every station generates 100,000 data packets.

The receiver has a sensitivity of –70 dBm. The transmitting power is 1 Watt. To guarantee a bit error rate of 0.001, the receiver receives only those data packets successful with a C/I-ratio of ≥ 13 dB. Data packets with a C/I-ratio < 13 dB are counted as collided. In the following scenario each station is connected with all other stations. So a collision happens if two stations send a data packet simultaneously.

To analyse the channel access mechanism in Figures 15, 16, 17, and 18 the throughput, the collision rate, the mean packet delay and the standard deviation of the mean packet delay are represented.

In Figure 15 is shown that to an input load of 50 % with the EY-NPMA-mechanism all data packets were transmitted. For higher input loads a synchronised sequence of data packets on the radio medium occurs. So the throughput commutes on a value of 48 %. That means, directly after a transmission a station begins with a new priority phase and the transmission medium is used ideally. The amount of data packets that really contend for the radio resource increases in proportion with the square of the input load. Thereby, as shown in Figure 16, in the area of lower loads it comes to a square rise of collisions. For higher loads the collision rate will be constant and reaches a value of 3.2 %.

To an input load of 40 % the mean packet delay reaches lower values (Figure 17). The intervals between the transmitted data packets are big enough, so that nearly all packets will be transmitted without delay. For input loads higher than 40 % too many data packets will be transmitted, so that many packets can be transmitted after several cycles. The mean packet delay of the EY-NPMA-mechanism converges to a value of 2500 µs. This limits results from the finite number of places on the stack. So, a limit transmission delay is ensured. This model leads to losses of data packets. This becomes noticeable in the difference of input load and throughput, because data packets which are not admitted in the stack will be deleted.

The standard deviation of the mean packet delay, shown in Figure 18, is low up to an input load of 25 %. To an input load of 50 % the standard deviation of the mean packet delay rises to high values of 8000 µs. For higher input loads the standard deviation of the mean packet delay remains constant.

For lower loads there are large transmission spaces between the transmitted data packets. In the EY-NPMA-mechanism no burst for the synchronisation will be transmitted. Because of this, a station which wants to transmit a data packet does not know if another station has begun with its access cycle. Therefore, stations with a data packet of high priority can reach the contention phase, although another station with a low priority is in an access cycle.

7 Evaluation of HIPERLAN on the problem of hidden stations

In this chapter the decentrally organisation of HIPERLAN with hidden nodes as asynchronised traffic sources will be analysed. Different scenarios with two overlapping cells each with eight stations will be considered. One station (hidden node) from a cell interferes with some stations of the other cell. In the following the cell with the hidden node will be called the interfering cell and the cell which will be interfered, the interfered cell.

For this analysis the same simulation parameters were used as in Chapter 6. The stations of the two cells are not connected with each other. So a collision can occur through stations from the interfering cell, without any direct radio contact to the interfered cell. This case happens when stations from the interfering cell send a signal simultaneously. These signals increase the noise level, which causes a C/I ratio lower than 13 dB. So, data packets can collide from stations without a direct radio contact to these stations.

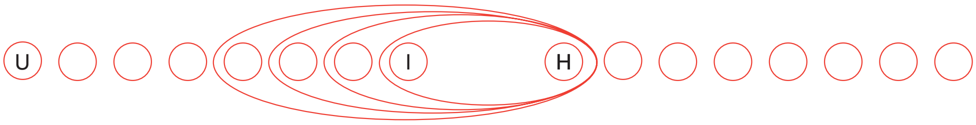


Figure 19 Scenario with one hidden station

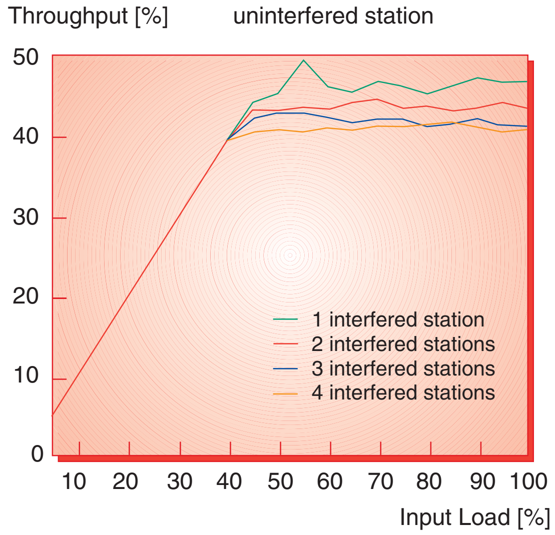


Figure 20 Throughput of the uninterfered station

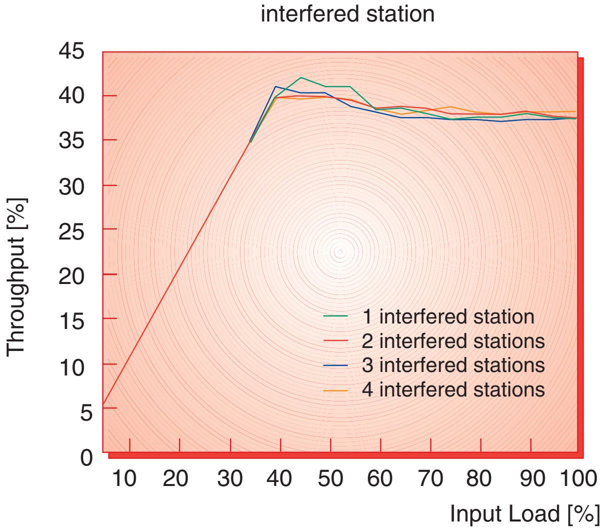


Figure 21 Throughput of the interfered station

To evaluate the channel access mechanism from the EY-NPMA-mechanism the throughput, the collision rate, the mean packet delay and the standard deviation of the mean packet delay will be analysed. These parameters will be shown for an interfered station and an uninterfered station.

The cell in this chapter will be interfered from one hidden node of the neighbouring cell. Different scenarios with one to four interfered stations from the hidden node (Figure 19) will be analysed. In Figure 19 the analysed interfered station is marked *I*, the uninterfered station *U* and the hidden node *H*. With the different scenarios with one hidden node the fundamental influence from hidden nodes to the decentral organisation of HIPERLAN can be analysed.

As foreseen the throughput from the uninterfered station decreases with an increasing number of interfered stations (Figure 20). The probability that an uninterfered station sends a data packet to an interfered station increases with the number of interfered stations. If an interfered station has a data packet with a smaller priority than the uninterfered station, the interfered stations do not send a burst in the priority phase. Then the hidden node considers the radio channel as free, at any time, the hidden node can begin with an access cycle and sends a signal. The interfered station receives this signal, and the data packet from the uninterfered station is collided. In this case the uninterfered station must retransmit its data packet and the throughput decreases.

The throughput of the interfered station, Figure 21, is independent of the number of interfered stations. The interfered station only sends a data packet if the hidden node does not interfere this station.

Figure 22 shows clearly that the collision rate of the uninterfered station increases

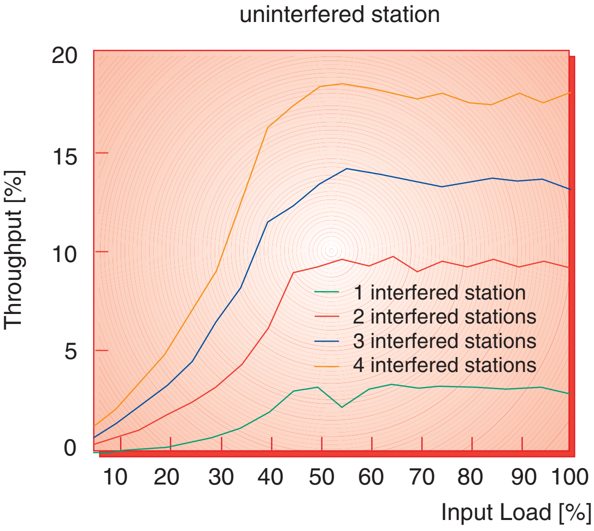


Figure 22 Collision rate of the uninterfered station

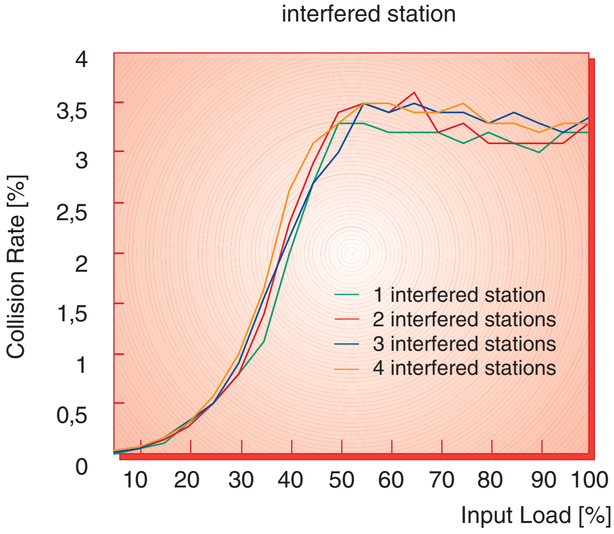


Figure 23 Collision rate of the interfered station

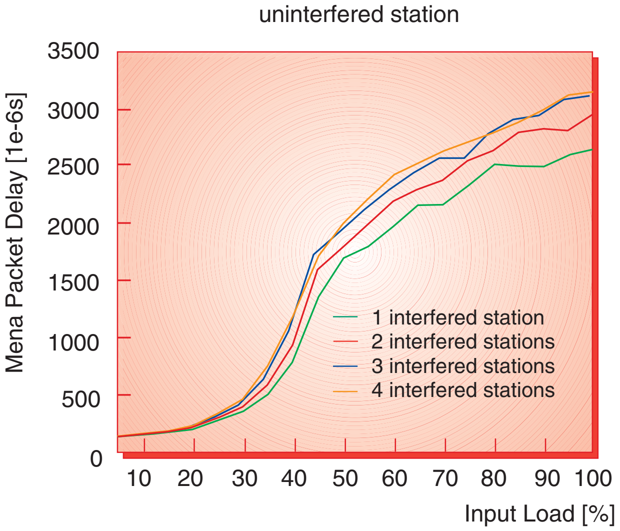


Figure 24 Mean packet delay of the uninterfered station

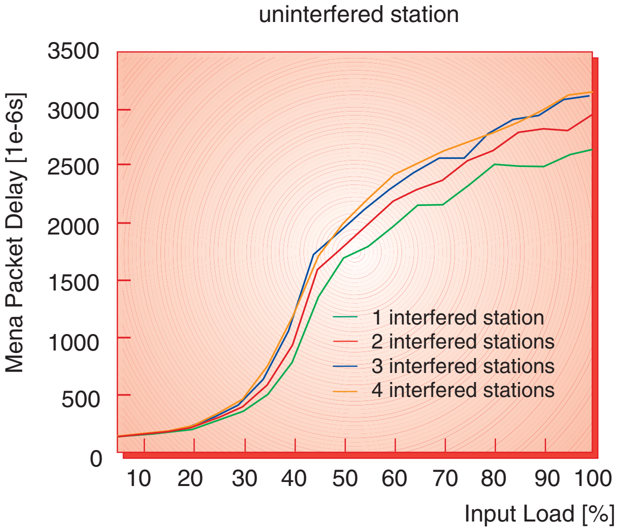


Figure 25 Mean packet delay of the interfered station

due to the increased number of interfered stations. Starting at 25 % input load the collision rate gets high values. For higher loads the collision rate reaches a constant value.

In spite of a linear increasing number of overlapped stations, the probability, that a data packet is directed to an interfered station, does not increase linear. For this reason the distance between the curves decreases with an increasing number of interfered stations.

The collision rate from the interfered station does not change with the number of interfered stations, as already described. Only in their own cell the stations cause a collision rate like in the one cell scenario. So the collision rate of these stations is very much lower than the ones of the uninterfered stations.

The curves from the mean packet delay in Figure 24 are independent of the number of stations. The mean packet delay of data packets of uninterfered stations increases only a little bit with an increased number of interfered stations, in spite of the increase of the collision rate. A retransmission of a data packet due to collisions has no influence on the mean packet delay, because the access cycles are very short.

The transmission delays are short up to an input load of 35 %. In this region the intervals between the access cycles are long enough, so that the interference by hidden nodes is slight. For lighter loads than 35 % the interference through the hidden nodes makes clear. So the mean

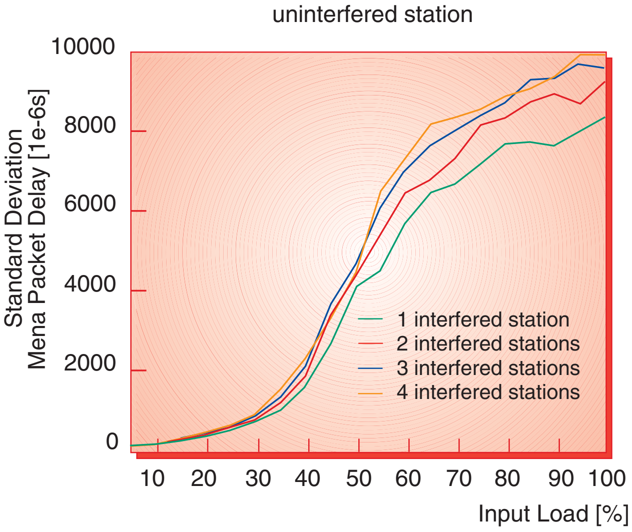


Figure 26 Standard deviation of the mean packet delay of the uninterfered station

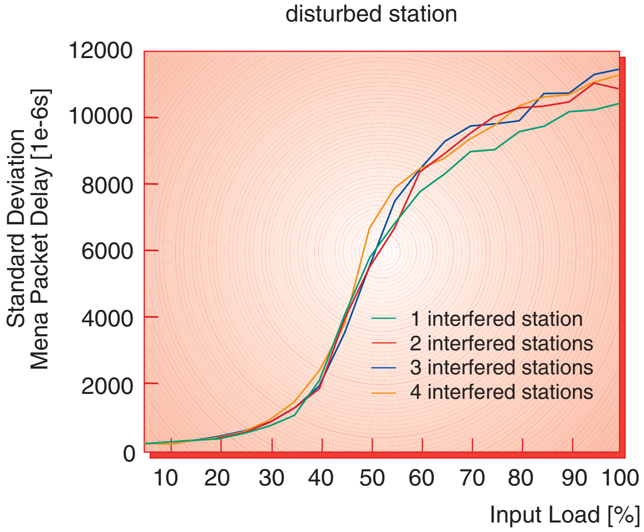


Figure 27 Standard deviation of the mean packet delay of the interfered station

packet delay increases end values from 2500 µs to 3000 µs.

The number of overlapped stations by the hidden nodes has no influence on the mean packet delay of interfered stations. The values are still low to an input load of 35 %. But the mean packet delay of these stations reaches higher values than the uninterfered stations.

The interfered station detects the radio channel at the beginning of the channel access cycle and in the priority phase. If the station detects a signal from the hidden node, the station finishes the access cycle. For this reason the mean packet delay increases.

For the standard deviation of the mean packet delay of the uninterfered station the same behaviour for the mean packet delay is given. Independent of the number of interfered stations the values are nearly identical (see Figure 26). Up to an input load of 35 % the standard deviation of the mean packet delay reaches low values. For higher loads dependent on the scenarios the standard deviation reaches end values between 800 µs and 10,000 µs.

Independent of the number of interfered stations the curves of the standard deviation of the mean packet delay of the interfered station shows nearly the same behaviour (see Figure 27). To an input load of 35 % the standard deviation of the mean packet delay of the interfered station reaches the same values like these from the uninterfered stations. For higher loads than 35 % the standard deviation of the mean packet delay increases stronger than at the uninterfered station. So the standard deviation already reaches high values for middle channel loads.

HMPDU	=	HIPERLAN MAC Protocol Data Unit
MSDU	=	MAC Service Data Unit
HMQoS	=	HIPERLAN MAC Quality of Service
EY-NPMA	=	Elimination Yield – Non-pre-emptive Priority Multiple Access

8 Conclusion

In this paper the structure of the HIPERLAN standard for wireless LANs was described. As one of the key parts, the channel access method, the EY-NPMA mechanism was evaluated by the analysis of the throughput, the collision rate, the mean packet delay and the standard deviation of the mean packet delay.

At first a one cell scenario with eight full connected stations was simulated. It turns out that the channel access mechanism is suitable for data, speech and video transmission [6]. In a next step the decentral organisation of the EY-NPMA mechanism with an additional asynchronous traffic of hidden nodes was analysed. It was noticeable that the data packets which were sent to the interfered stations, were transmitted with a considerable collision rate. Already for lower input loads the collision rate from these data packets reached high values. Nevertheless, the EY-NPMA mechanism is suitable for data, speech and video transmission in the case of overlapped cells.

References

1 Herrig, W. *Simulation tool for the evaluation of access protocols for RLANs*. Aachen, Institut für Kommunikationsnetze, March 1994. (Diploma thesis.)

2 Hong-Yon Lach. *EY-NPMA ver. 0.2*. 1994. (ETSI RES 10.)

3 Phipps, T. HIPERLAN medium access control. In: *IEEE/ICCC/WCN/PIMRC '94,* 1994, 870–874.

4 Herrig, W, Plenge, C. Evaluation from the decentral channel access mechanism of HIPERLAN. In: *ITG conference Ulm.* Aachen, Institut für Kommunikationsnetze, September 1994, 371–379.

5 ETSI. *Radio equipment and systems (RES) : HIPERLAN : functional specification*. Draft. 1995. (ETSI DE/RES 10-01.)

6 ITU-R. *Radio local area networks (RLANs)*. ITU, Radiocommunication Study Groups, September 1994. (Doc. 9C/TEMP/30-E.)

Code Division Multiple Access – hot topic in mobile communications

BY JAN ERIKSEN AND OLE DAG SVEBAK

1 Introduction

During the last decade first generation land mobile telecommunications systems such as NMT, TACS and others have spurred a new era in the world of telecommunications. Today, second generation systems such as GSM, DECT, ERMES and others are being introduced around the world. These are digital systems tailored to specific application areas, such as cellular applications, wireless office communications or paging. However, in order to meet expected market demands and provide wireless telecommunications services to "anyone, anywhere, any time" work has already been progressing in research institutes and standardisation bodies around the world for a number of years. These efforts aim at extending the service repertoire provided to the user beyond that of earlier systems, as well as integrating the application areas of second generation systems into one telecommunications system, in Europe termed the Universal Mobile Telecommunications System, UMTS (refer to article on UMTS elsewhere in this issue of Telektronikk).

One of the many key features of a mobile telecommunications system is the multiple access technique. The multiple access technique may be considered as the way in which the radio interface is structured in order to facilitate simultaneous access of a large number of mobile users to the telecommunications system. The large interest in the multiple access technique is justified by the fact that the technique may strongly impact the cost of the radio sub-system, the number of users to be supported at any one time, etc.

The radio interfaces of first generation systems are based on frequency division multiple access (FDMA), whereas second generation systems generally utilise time division multiple access (TDMA). Following the success of first and second generation systems as well as recent advances in signal processing technology, the last decade has seen a rapidly growing interest in direct sequence code division multiple access (DS CDMA) as a candidate multiple access technique for both the land mobile as well as the satellite component of third generation systems.

DS CDMA may be considered as an application of the direct sequence spread spectrum communication technique to multiple access applications. Traditionally, direct sequence spread spectrum has been widely used in military radio applications. This is largely due to the favourable anti-jamming capabilities of the spread spectrum signal, as well as the high degree of privacy provided by the secrecy of the spreading code used on a particular radio link. This makes the signal difficult to detect and decode for third parties (intruders), as well as hard to jam, thus facilitating a reliable and private means of communicating in a hostile environment. In a multiple access scenario, the radio environment is usually characterised by noise and interference from other users of the system, and it is therefore primarily the anti-jamming capabilities of the spread spectrum technique which are sought exploited. Privacy may be provided using suitable encryption schemes, rather than keeping spreading sequences unknown to third parties. In particular, research interest has been focused on DS CDMA, in which the user information is multiplied by a pseudo-noise sequence prior to transmission, thereby spreading the information signal in the frequency domain. In the receiver, a corresponding correlation operation with the same PN sequence returns the user information.

The remainder of this article is structured as follows. In section **two** the basic principles of direct sequence spread spectrum and its application to multiple access are described. Section **three** gives an overview of some decoding algorithms and possible receiver structures for DS CDMA multi-user and single-user receivers, e.g. base station and mobile station receivers. We have attempted to keep the description at a non-mathematical level, emphasising the capabilities and limitations of the various decoding schemes in a multiple access environment, particularly with regard to interference performance and computing power requirements. In section **four** a brief description of the Qualcom CDMA system is given with a highlight on the multipoint (base) receiver section. Section **five** aims at high-lighting some of the pros and cons of DS CDMA. In this context we have also attempted to emphasise some of the main differences between DS CDMA and TDMA. Finally, section **six** briefly summarises the article.

2 Principles of DS CDMA

Generally speaking a CDMA signal is generated by *transforming* the low frequency information signal or the high frequency IF/RF-signal by means of applying a *coded* signal to it. This transformation can be done in several ways, but we often distinguish between two major principles of CDMA. First and most referred to we have Direct Sequence or DS CDMA which shall be discussed further in this section, second we have Frequency Hopping or FH CDMA which will be described only briefly in the next paragraph.

As the name indicates FH CDMA implies transmission of a number of different frequency components. The hopping pattern is determined by channel type, mobile subscriber number, and other parameters connected with the communication. FH CDMA is even diverted into two different principles. First we have **fast** FH where several carriers are transmitted during the period of the information symbol, and second we have **slow** FH where changing of carrier happens only every information symbol or even every few information symbols.

In DS CDMA the information signal is transformed into the spread signal by applying a spreading code. The transformation is a logical or mathematical operation on the baseband signal or even on the intermediate-/high-frequency signal.

2.1 Basic building blocks of DS CDMA

In case of bipolar signals, spreading is done by a multiplication between the coded information bits d_c and the spread-

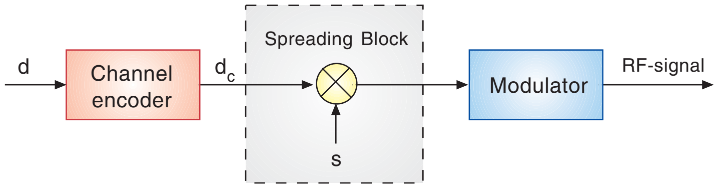


Figure 2.1 DS CDMA transmitter including channel encoding, spreading function and modulator

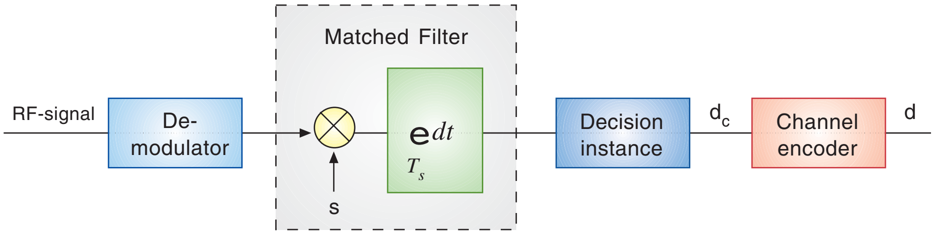


Figure 2.2 Simple matched filter receiver with demodulator, matched filter, decision instance and channel decoding block

ing sequence *s*, Figure 2.1. Different types of coding or encryption are optional.

In most practical systems the rate of the spreading sequence *s* (chip rate), is much higher than the rate of the information signal d_c. The ratio between these, often referred to as processing gain P_G, is normally a factor of a few tens or a few hundreds. The chip-stream from the spreading block is fed into the RF-modulator. All of these blocks will be implemented in one or a few integrated circuits.

Figure 2.1 shows the transmitter part of one user or channel. In the multipoint transmitter (base station) the signals intended for the different mobiles will be combined, either at base band level or at the IF/RF signal.

The DS CDMA receiver performs the complementary operations of the transmitter. Figure 2.2 shows a simple matched filter receiver consisting of a demodulator, despreading block, decision instance and a channel decoding block. Despreading is done by multiplying the discrete quantised signal from the RF-front-end with a replica of the spreading code used in the transmitter, together with an integration over the symbol interval, T_S. The decision instance will in this simple receiver decide whether the input is higher or lower than a given threshold. In a later chapter we shall see that the multiple access receiver can be far more complex than this, and it has a far more exhausting task to solve.

The optional channel encoding and decoding block is of the same type which would be used in a conventional system, including error correction coding and block interleaving.

Table 2.1 Code length, family size, non-peak auto-correlation values, and variance of cross-correlations given for m-sequences, Gold codes, and Walsh functions

Family	Code length L	Family size	Auto-corr non-peak value[1]	Auto-corr non-peak standard deviation	Cross-corr peak value[1]	Cross-corr standard deviation
m - sequences	31	6	1	1	9...11[2]	5.7
	127	18	1	1	17...41[2]	11.3
	511	48	1	1	33...113[2]	22.6
Gold codes	31	33	9	5.7	9	5.5
	127	129	17	11.3	17	11.2
	511	513	33	22.6	33	22.6
Walsh codes	32	32	32	17.4	32	4.8
	128	128	128	53.0	128	10.3
	512	512	512	161.0	512	21.5

[1] this value may be negative or positive
[2] lower number valid for preferred pairs

2.2 Spreading codes

The spreading codes are most important for overall system performance, especially the codes used in the reverse link (from mobile to base station). The most important feature of spreading codes is probably their cross-correlation. The cross-correlation between code words is vital with respect to the ability of the base to separate signals from different mobiles using different spreading codes. The cross-correlation should be lowest possible for all *t*. A low cross-correlation implies that the output from the matched filter corresponds to that particular code word in the received signal and not other interfering code words, e.g. better multiple access properties. It seems that most families of spreading-codes have highest cross-correlation when $\tau > 0$. Another important aspect of the codes is their auto-correlation, which determines a receiver's ability to lock on a received code when being synchronised with it. The important issue is the ratio between the non-peak auto-correlation value ($\tau > 0$) and the peak auto-correlation value ($\tau = 0$), which normalised equals one.

Equations 2.1 and 2.2 give the expressions for auto-correlation $\phi_{kk}(\tau)$ and the cross-correlation $\phi_{kl}(\tau)$.

$$\phi_{kk}(\tau) = \sum_{i=1}^{L} S_k(i) S_k(i+\tau) \qquad (2.1)$$

$$\phi_{kl}(\tau) = \sum_{i=1}^{L} S_k(i) S_l(i+\tau) \qquad (2.2)$$

These equations are valid if the spreading sequences are antipodal, e.g. the chip values are +1 and –1. Using logical representation a match between chips should give positive contribution and a mismatch negative.

The spreading codes should even be well balanced, e.g. there should be equal numbers of ones and zeros (binary code) and groups of consecutive ones and zeros should be approximate the same. The imbalance of the code word should be at most one, e.g.

$$|\textit{number of ones} - \textit{number of zeros}| < 1.$$

The balance of the code have implications on the frequency spectrum of the transmitter.

The cross-correlation between code words will often take a number of discrete values, dependent on code type and offset in time. Used in an asynchronous manner the standard deviation of cross-correlations between codes in a family will often give a better indication of the actual size than the discrete values or limits which cross-correlations will have or lie within. The values as given in Table 2.1 have been calculated by means of Matlab and C-programs. In literature the correlation values are often given in one of two different forms, non-normalised or normalised. In the non-normalised form the correlation value is dependent on the code length and its absolute maximum value is the code length *L*. In the normalised form the correlation value is normalised with the code length *L* and the absolute maximum value equals one. Correlation values in Table 2.1 are non-normalised.

The columns number four and five need further explanation. These columns give maximum auto-correlation values for τ more than zero, e.g. the value when τ equals zero is excluded from the calculation. Therefore, these columns indicate the ability of the codes to lock on an incoming signal spread with the same code in the acquisition mode.

There are several families of codes which have different properties concerning code length, alphabet size (size or number of levels a code bit may have), family size (number of code words which naturally belong to one group), auto- and cross-correlations, Table 2.1. Some of the well known codes are Maximum Length sequences [1], Gold codes [1], Kazami codes [1], and Hadamard [1] or Walsh codes (functions). All of these codes are binary and linear, and they can be generated by using shift registers with feedback.

The maximum length codes (m-sequences) are probably the best known type of codes. A shift register with *l* bits can be used to generate a sequence with a *maximum length* of $L = 2^l - 1$ before the sequence repeats. Used in a synchronous manner the auto- and cross-correlation properties of m-sequences are excellent, but when used asynchronously the cross-correlation between code words increases. In Table 2.1 there are two numbers in the cross-correlation column. M-sequences which have cross-correlations equal to the lower number are called *preferred pairs*. The major disadvantage of m-sequences are the limited number of code words of given length *L*.

The often used Gold codes are generated by combining two *preferred* maximum length sequences with a chip-offset between the two. The Gold code has an alphabet size of 2, a code length of *L* and a family size of *L* + 2, e.g. while there are rather few maximum length sequences of length *L*, a pair of maximum length sequences can be used to generate *L* + 2 Gold codes. The non-peak auto-correlation and peak cross-correlations of Gold codes are similar. Using different maximum length sequences generates a new group of Gold codes. Code words from 'different' Gold code groups have bad correlation properties, even when synchronised.

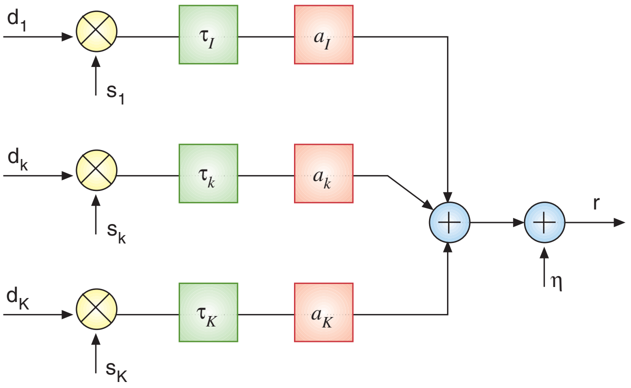


Figure 2.3 Multiple access model with K mobiles, d, s, a, and t represents user data, spreading code, signal amplitudes and relative delays. h is the additive Gaussian noise in the receiver

The Walsh functions or Walsh codes are another often used family of codes. These codes have a code length and a family size of $L = 2^N$ where *N* is a positive integer. When synchronous, these codes are perfectly orthogonal, e.g. they have cross-correlation zero, but when asynchronous their cross-correlation is very much dependent on the particular pair of codes used, some will have cross-correlation zero while others will have a very high correlation. Even the non-peak auto-correlation of Walsh codes is dependent on the actual code word. Some code words will have very high non-peak auto-correlation values while others may have lower values, but never zero. Walsh functions are generated according to the following formula:

$$H_N = \begin{bmatrix} H_{N-1} & H_{N-1} \\ H_{N-1} & \overline{H_{N-1}} \end{bmatrix} \tag{2.3}$$

where $H_0 = 0$

In the IS-95 standard quite a different type of spreading sequences are used. Here, the sequence used during each modulating symbol represents only a subset of a much longer code, which is a maximum length sequence. If the length *N* of the code is much longer than the length *W* of the subset, the correlation between such subsets have a normal distribution with a standard deviation close to the square root of *W*.

2.4 Synchronous and asynchronous DS CDMA systems

Normally, the term synchronous CDMA reception indicates that the modulating symbols received from every mobile are synchronised at the base station receiver within one chip. If the symbols arrive at the receiver with higher offset in time than the duration of one chip they are called asynchronous.

Figure 2.3 shows the simplified model of the received signal *r* from *K* users. d_k is the generated bit (included error control etc.), s_k is the spreading code, τ_k ($0 < \tau_k$ < modulation symbol period) is the relative signal delay due to different distances between mobiles and base and a_k is the signal amplitudes due to path loss or transmitted power. η is the additive Gaussian noise in the receiver. Of course *d, s*, τ, *a*, *r*, and η vary with time.

A more complete and accurate description of the received signal is given in equation (2.4).

$$r(t) = \sum_{l=-\infty}^{\infty} \sum_{k} a_k(t) d_k(t - iT - \tau_k) \cdot s_k(t - iT - \tau_k) \cdot \cos(\omega_0 t + \theta_k) + \eta(t) \tag{2.4}$$

where subscript *k* denotes user *k* and

a_k is the signal amplitude

d_k is the (*i*'th) data bit

τ_k is the relative delay of the symbol

s_k is the spreading code

ω_0 is the carrier frequency

θ_k is the random phase of the RF-signal

η is the additive Gaussian noise.

The reception of many signals in the base station receiver and simultaneous estimation of bits from several users is a problem which has been studied intensively. Different approaches to solve different receiving tasks are discussed in chapter 3.

3 DS CDMA receivers

Traditionally, the term receiver has meant an RF/IF to LF conversion obtained by using analogue components in analogue systems or together with a few digital components in digital systems. In modern digital systems most of the transformation from RF/IF to information bits can be and often is done in one integrated digital circuit only, immediately after an analogue to digital conversion close to the RF front-end itself. The appropriate term of the demodulating/decoding process would therefore be a *receiving algorithm*. In spite of this the term receiver is used in this article, with few exceptions.

The DS CDMA receiver can either be the very simple *matched filter* type, where decision on received information bit is done immediately after the matched filter, or it can be of a *joint detection* type. The matched filter receiver is probably the most commonly used, but this section will concentrate on the complex joint detection receiver. The term joint detection means that the receiver algorithm use all known information about **all** users in the process of decoding **one** user.

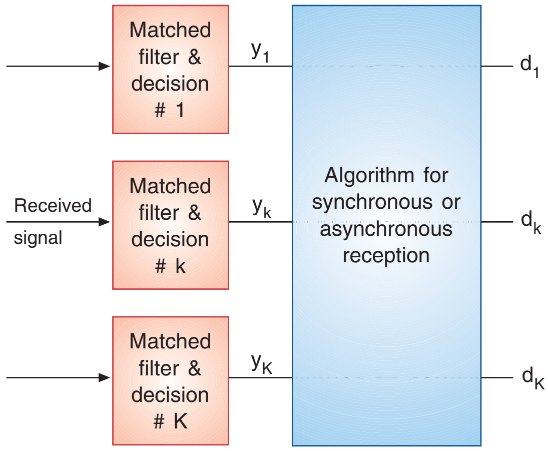

Matched filter & decision # 1
y1
d1
Received signal
Matched filter & decision # k
yk
Algorithm for synchronous or asynchronous reception
dk
Matched filter & decision # K
yK
dK

Figure 3.1 The general multiple access receiver, consisting of matched filters and the decision algorithm

In a DS CDMA system the tasks which mobile- and base receiver are given are a little different. The mobile shall extract one specific code from a signal which in some meaning is combined from a set of codes. The base is given quite a heavier task, which is to demodulate many users when the received signal is a sum of signals sent by the different mobiles, with variations in amplitude and relative timing and multipath propagation.

Before discussing different receiver strategies a simple description of the complex propagation environment will be given.

3.1 The cellular propagation environment

In a cellular system the DS CDMA receiver at the multi-point (base) has two major problems to solve. One is the so-called *near-far* problem, which implies that signals received from mobiles close to the base tend to mask signals received from distant mobiles, when the transmitted power from the mobiles is the same. The near-far problem is of course closely linked to the non-perfect orthogonality between the spreading codes. Traditionally, this has been solved by using strict power control at the mobiles, but lately, the use of complex reception algorithms have solved this problem, at least to some degree. Power control will, however, be used in the mobiles in most practical systems, both as an interference decreasing measure and as a power saving measure.

The other problem is the multipath propagation properties of the radio channel. The mobile will typically have an omni-directional antenna, and the base will have an antenna with a given antenna diagram. Therefore, reflecting objects such as buildings, hills or mountains will cause multiple paths between mobile and base, every path having a different time delay, amplitude and phase which give a channel impulse response of the following form (see article on multipath propagation in this issue of *Telektronikk*):

$$\sum_{k=1}^{K} a_k \delta(t-\tau_k)\exp(j\theta_k) \tag{3.1}$$

where subscript k denotes impulse k and

a_k is the amplitude of the impulse response

τ_k is the relative time delay of the impulse response

θ_k is the phase of the impulse response.

If the delay is long compared to the symbol period and the delayed signals are strong compared to the direct path, then these multiple paths will cause severe intersymbol interference which has to be resolved both in the mobile- and the base station receivers.

3.2 Single user receivers

In most practical systems the base will transmit signals intended for all mobiles in a synchronous manner. This implies that the mobile receiver always receive synchronised signals. Even the amplitudes of the different user signals transmitted from one base are most likely equal, though a system which use increasing signal strengths with increasing distance to mobile is possible.

Modern systems will most likely be designed to allow multiple data rates assigned to different users, or the possibility to allow dynamic capacity to some users if necessary. This can be done by assigning several traffic channels to one user.

In this case the mobile will use an algorithm which is optimised for detection of only a subset of all codes in the received signal. Though a maximum likelihood algorithm has an exponential growth of complexity with number of codes, a low number of codes means that such an algorithm can be used with a significant improvement in performance. Single user receivers can be found in [2].

3.3 Multi user receivers

The base must be able to demodulate a large number of signals which are added together in a synchronous or asynchronous manner. This can put exhausting demands on the algorithm in the base. Receivers for multiple access are becoming very complex and their performance is close to the optimum receiver, which even implies a performance which is close to a single user receiver in a single user environment. Multiple user receivers can be found in [3].

The optimum receiver is a maximum likelihood algorithm which examines every combination of the variables in the received signal and after this examination chooses the most likely path, e.g. the most likely estimation of signal amplitudes *a*, relative time delays τ and transmitted data bits *d*. The optimum receiver has a complexity which is exponential with the number of users. So far, this vast complexity has meant that a practical solution using such an algorithm has been impossible.

So, instead of exploring the optimum receiver, a lot of work has been done in looking at the *optimum linear receiver*. As its name implies, the algorithm has a complexity which is linear with the number of users and it has a performance which is similar to the optimum receiver and far better than the simple matched filter receiver.

A large number of linear receiver approaches has been investigated. One possible algorithm is a decorrelating algorithm [4] and [5]. This algorithm simply multiplies the received signal with a "decorrelating matrix" which resolves the interference caused by cross-correlation between spreading codes. A major problem with this approach is that an inversion of a cross-correlation matrix is necessary to obtain the "decorrelating matrix", and this is an exhausting operation, especially in the asynchronous case.

Other often used algorithms are interference cancellation algorithms [6]. These algorithms estimate the signal strength of the strongest interferer in the received signal. Then, this knowledge together with an estimation of the information bit is used to subtract (cancel) this interferer from the received signal. The same operation is then done for the next strongest interferer. The cancelling of interfering signals is often done in several stages, e.g. when every estimated interferer has been removed from the received signal the process can be repeated on parts of or even the complete received signal. Using a different approach together with the knowledge of the estimated interference strengths and information bits a better estimation and thereby a better cancellation and decision can be made.

Different decoding algorithms have different qualities when used in different system environments. One algorithm will be preferred in an urban environment with high data rate and short multipath delays, and another will be preferred in a rural area with lower data rates and long multipath delays (delays which are much longer than the data- or symbol period).

3.4 The RAKE receiver

The RAKE receiver is the classical single-user receiver used to combat multipath effects usually experienced in direct sequence spread spectrum communications. A number of papers have discussed various aspects of the RAKE receiver. The probably most informative description is given by Turin [7]. In this section we attempt to briefly summarise the principles of operation of the RAKE receiver, also pointing out the most important features impacting the performance of this type of receiver.

3.4.1 Principles of operation

In order to describe the principles of operation of the RAKE receiver, it is useful to start from an ordinary matched filter detector for binary signalling, designed to detect the information transmitted by one user only, as illustrated in Figure 3.2.

The matched filters correlate the received signal with a replica of the spreading code. Whenever a symbol match occurs in the matched filter, a correlation peak is output from the envelope detector. For binary antipodal signalling (e.g. BPSK) this corresponds to positive and negative correlation peaks (positive and negative amplitudes). At the decision instant, the decision circuit outputs the symbol (0, 1) corresponding to the most likely transmitted symbol. For a binary signalling scheme the decision circuit is essentially a threshold detector, whose output depends on whether the input amplitude wave form exceeds the threshold level or not at the sampling instant.

The matched filter receiver performs very well in environments where the disturbances are characterised as additive white Gaussian noise. In cellular systems the signal is often disturbed by intersymbol interference (due to multipath reflections) or interference from other users. The matched filter receiver illustrated in Figure 3.2 does not account for this. However, in mobile applications of DS CDMA the bandwidth of the signal may be in the order of MHz, as dictated by the user information symbol rate and the chip rate. As the bandwidth of the receiver increases it becomes possible for the receiver to resolve individual incoming replicas of the transmitted signal, caused by multipath transmission (refer to the introductory paragraph to this section).

Usually, the time resolution of the receiver is equal to the PN sequence chip length. (In typical mobile applications this would be in the order of a micro-second or less.) Provided that the delays between incoming replicas of the transmitted signal is greater that the duration of the PN sequence chip length, the receiver is able to resolve the individual replicas arriving at different instants in time. This facilitates what is often termed time diversity reception, which indeed forms the basis for the operation of the RAKE receiver.

Turning to the simple matched filter detector of Figure 3.2, it is clear that, if replicas of the same symbol (due to multipath) arrive with relative delays longer than the PN sequence chip duration, multiple correlation peaks will be observed at the output of the envelope detector. Thus, for one symbol multiple correlation peaks may be observed. The objective of the RAKE receiver is to combine such multiple replicas into one distinct symbol wave form as originated from one user, whilst attempting to suppress all other received symbols and users.

In order to combine incoming replicas of the same transmitted symbol in the receiver, very accurate estimates of the propagation channel must be available to the receiver. In the RAKE receiver a channel sounding receiver may be used to obtain estimates of amplitudes and relative time delays of the incoming replicas. (It is worth noting that the rapid phase fluctuations of the mobile channel usually exclude phase shifts estimates, thus non-coherent reception may be assumed.) These estimates are used to activate the taps of a finite impulse response filter at the envelope detector output, as illustrated in Figure 3.3.

Incidentally, the tooth-like structure of the finite impulse response filter at the

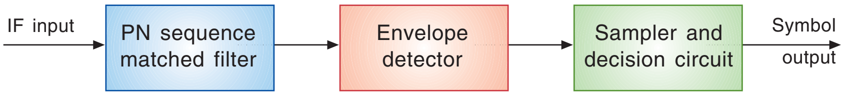


Figure 3.2 Single user matched filter detector for binary symbol alphabet

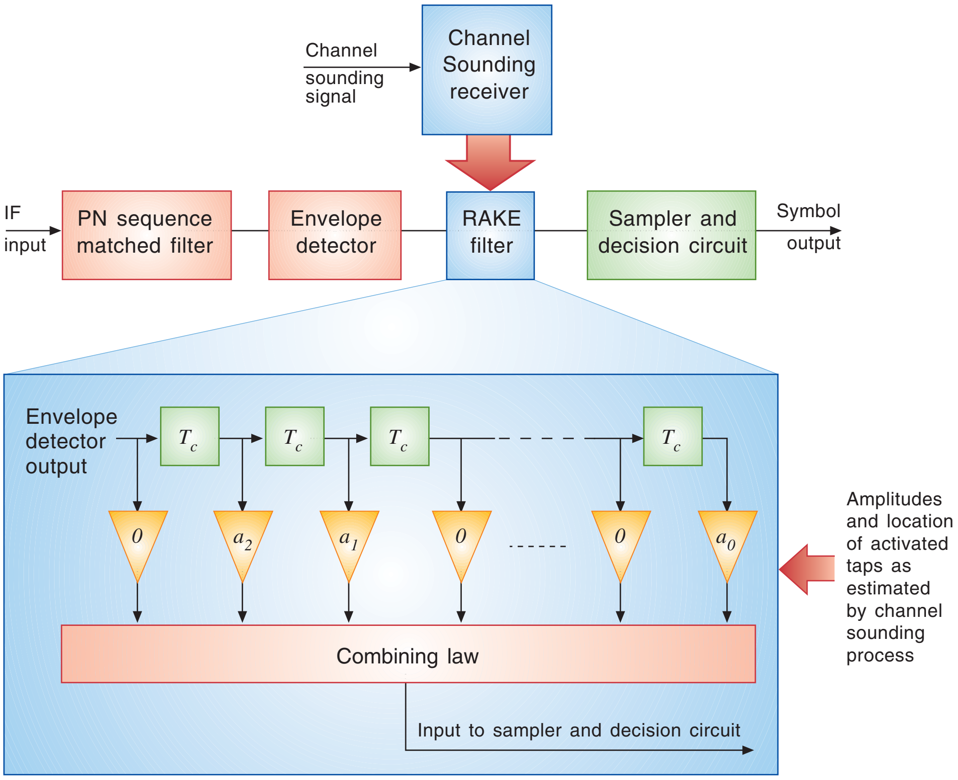

Channel sounding signal
Channel Sounding receiver
IF input
PN sequence matched filter
Envelope detector
RAKE filter
Sampler and decision circuit
Symbol output
Envelope detector output
T_c
T_c
T_c
T_c
0
a_2
a_1
0
0
a_0
Amplitudes and location of activated taps as estimated by channel sounding process
Combining law
Input to sampler and decision circuit

Figure 3.3 Structure of RAKE receiver

envelope detector output has indeed given the name to this type of receiver, as was first coined by Price and Green [8]. A few points are worth noting about the RAKE receiver:

- The RAKE filter ideally reverses the impact of the propagation channel on the received signal. It may therefore be interpreted as a filter matched to the propagation channel impulse response. (The matched filter prior to the envelope detector is matched to the transmitted signal, not to the propagation channel.) The time separation between adjacent RAKE filter taps is usually equal to the PN sequence chip length.
- The RAKE combining law, e.g. the way in which the delayed replicas of the received signal are being combined (e.g. summed, as illustrated above) at the decision instant, is an interesting feature impacting the overall performance of the RAKE receiver. The optimal combining law depends on the statistical properties of the received replicas of the transmitted signal. Whatever the characteristics of the optimal combining law, its objective is to emphasise credible data relative to less credible data.
- The channel sounding process should result in very accurate estimates of the propagation channel impulse response (amplitudes and relative delays). If this is the case, only a limited number of taps of the RAKE filter will be activated at a given instant of time, causing the RAKE filter to output a single strong peak corresponding to the one symbol received multiple times. Thus, it should be clear that the accuracy of the channel sounding process is vital to the error performance of the receiver.
- The sounding receiver may in itself be implemented as a finite impulse response filter, matched to the sounding signal. The sounding signal may for instance be a linear frequency sweep or a PN sequence, which both have good auto-correlation properties, if well chosen. Care needs to be taken so as to ensure that the sounding signal is well received, e.g. by allocating a separate physical channel for the sounding signal or by increasing the power of the sounding signal. Propagation channel estimation may also be facilitated by transmitting a signal sequence known to the receiver within the user information bit stream, as in GSM, allowing the receiver to estimate the propagation channel impulse response from the received signal sequence. Whatever technique is chosen, the resulting spectrum efficiency and battery power requirements, which are real constraints in mobile applications, must be traded off against the expected improvement in receiver performance.

3.4.2 Receiver performance

The mobile radio environment may usually be characterised by significant signal variations over short distances and significant time dispersion of the transmitted signal as observed in a receiver. In such a context it is therefore particularly important to study the effects of interference on the receiver performance. This includes interference from other users of the same frequency band, as well as inter-symbol interference caused by the multipath propagation channel.

Considering first interference from other users; provided that the PN sequences have low cross-correlation between them and the interfering signal(s) are not much stronger than the desired signal, the correlation performed by the matched filters prior to the envelope detectors will tend to despread the desired signal and further spread all other signals. Thus, interfering users will ideally appear as low-level noise at the envelope detector output, whereas the desired signal is retrieved. Secondly, concerning inter-symbol interference between successive symbols on the desired link (due to multipath effects) it should be noted that it is highly unlikely that successive symbols will occur at taps which are activated at the decision instant. Therefore, the RAKE filter will tend to act as an inter-symbol interference canceller. The interference cancelling properties of the RAKE receiver may be improved by increasing the number of taps in the RAKE filter.

The receiver performance with respect to error probability is governed by a number of factors. It is possible to improve the performance by increasing the number of taps in the RAKE filters, that is, to let the RAKE filters span a longer time period. The performance may also be improved by decreasing the (time) spacing between adjacent taps in the RAKE filter. However, the time resolu-

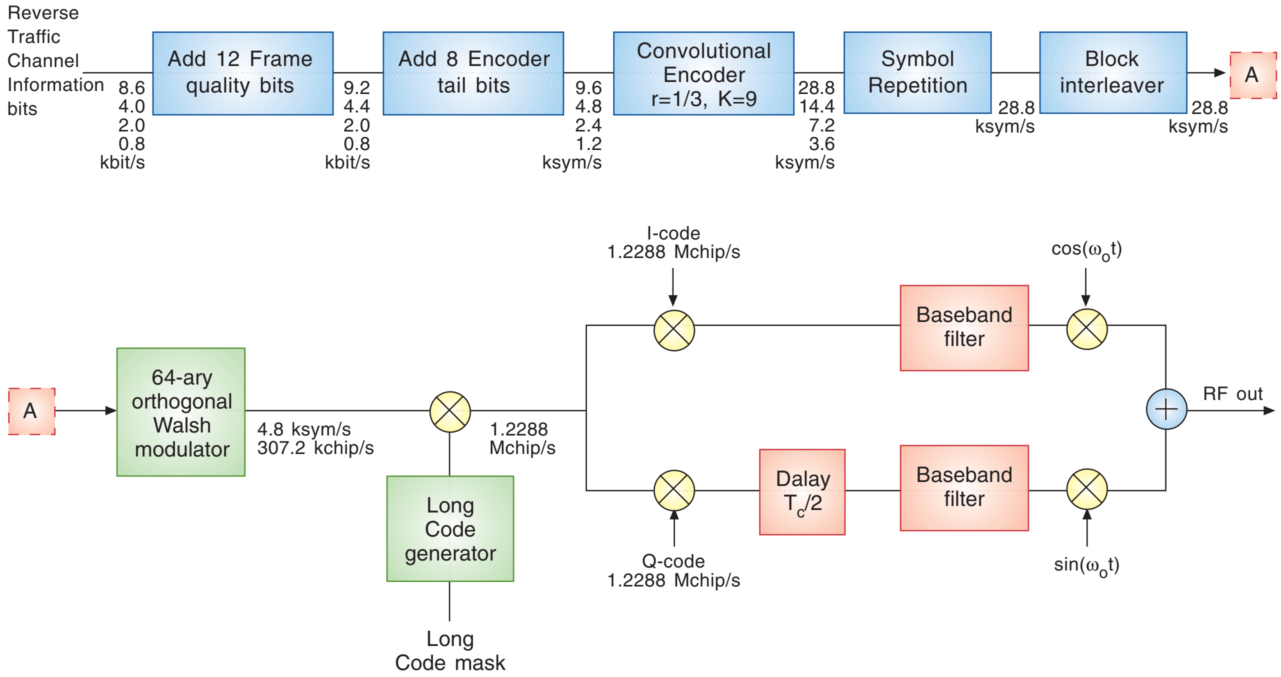


Figure 4.1 Building blocks used for traffic channel in the IS95 mobile transmitter

tion of the RAKE filter is determined by the PN sequence chip length. It should also be noted that the receiver performance is strongly influenced by the accuracy of the propagation channel estimates resulting from the channel sounding process. From this it should be clear that performance improvements are possible, at the expense of increasing the signal bandwidth, increasing sounding signal power or increasing the complexity of the receiver.

Detailed error analysis of the RAKE receiver becomes mathematically intractable as the propagation channel model approaches the real world. A number of simplifications and assumptions are therefore usually made in order to obtain performance estimates. Error performance aspects are extensively dealt with in the literature, e.g. [7], [9], and [10].

4 Brief description of the Qualcomm CDMA system

Most existing CDMA systems today are personal cordless telephone systems and not cellular systems. The processing gain is often very high and number of simultaneous users much less than the code length. These systems are mostly used in cities, where multipath delays are shorter than the symbol period and often even shorter than the chip period.

The Qualcomm system, which complies with the IS95 standard [11], is a cellular system and today probably the most known CDMA system. This system is far more complex than a simple “multiplication with a spreading code”, and it is beyond the scope of this article to explain all details of the system, but a few major characteristics will be given.

The maximum user data rate is 8600 bit/s. The radio frequency bandwidth is approximately 1.23 MHz. The system offers 64 logical channels. At the forward link this includes 1 pilot channel, 1 synchronising channel, 7 paging channels, and 55 traffic channels. The sync and paging channels can be replaced by traffic channels one by one. At the reverse link there are 9 access channels and 55 traffic channels.

Unfortunately, IS95 gives detailed information only on the transmitter part of mobile and base. The specification only states that the receivers should do the complementary operations of the transmitter. A brief description of the more complex transmitter of the mobile will be given, Figure 4.1.

Information bits are generated in the voice encoder or at the data connector input. To these 12 frame quality bits and 8 encoder tail bits are added. At this point the bit rate is 9600 bit/s. Then a rate 1/3 convolutional encoder with constrained length 9 is used. After convolutional encoding a block interleaver with a length of 576 bits (20 ms) is used. These bits are used six and six to choose between 64 different Walsh functions, generated as shown in section 3.

After the Walsh encoder three different spreading codes are used. The code which gives diversity between mobiles is of length $2^{42} – 1$. In fact, all mobiles are using the same code, but depending on mobile subscriber number, channel type, and channel number each mobile is using the code at a different time- or chip offset than any other mobile. In theory, two mobiles could be using the code with only one chip relative offset, but with the vast length of the code the probability of occurrence is very low. However, only a subsection of this long code is used every modulating symbol.

Then two different codes of length 2^{15} are used, one at the in-phase branch (I-code) and one at the quadrature branch (Q-code). These codes are the same for every mobile. The signal at the Q branch is delayed with 1/2 chip period.

Finally, the I and Q signals are filtered by means of a high order filter, which is approx. 620 kHz wide, before the two branches are multiplied by a 0 deg and 90 deg carrier respectively, and added.

The system incorporates power control at the mobiles, both an open loop control which adjusts the output power according to measured input power, and a closed loop control which adjusts output power in response to commands from the base station.

The mobile incorporates a four tap RAKE- receiver as stated in the specification (se RAKE-receiver section), where 3 taps are used to combine the signal from one base and the last tap to look for bases which give better reception.

5 Pros and cons of DS CDMA

The objective of this section is to identify some of the main pros and cons of DS CDMA applied to a cellular system. These considerations are of a general nature, and aims at highlighting some potential concerns and advantages of DS CDMA as an access scheme, compared to other access schemes. The issues addressed below are covered in the literature, e.g. in [12], [13], [14], [15] and [16].

5.1 Spectrum efficiency and capacity

Spectrum efficiency is considered one of the most important parameters of a cellular system, yet it is hard to define precisely what it actually is. However, in general terms spectrum efficiency may be considered as the radio capacity available for a given cost.

For a cellular system this translates into the number of active users or amount of traffic which can be accommodated within a given geographical area and frequency band, given assumptions on service quality (bit error rate performance, outage probability, etc.), number of cells within the area (and hence financial cost) and transmission parameters (assumptions on power control, voice activity detection and discontinuous transmission, etc.). Moreover, some assumptions will only apply to one multiple access technique, thus further complicating a fair comparison between different multiple access techniques.

Given these disclaimers it seems from the literature available (e.g. [12]–[16]) that there is a general tendency to favour DS CDMA to second generation cellular TDMA systems as far as capacity and spectrum efficiency are concerned. In fact, some papers (e.g. [14]) claim that DS CDMA represents a quantum increase in current cellular capacity, with no other proposed scheme even approaching the performance of DS CDMA. [15] indicates that DS CDMA is a promising candidate for satellite applications as well. Others (e.g. [12]) are more reluctant to claim that DS CDMA by far outperforms optimised second generation systems.

A simple analysis can be carried out to be able to determine the parameters involved. For a given modulation technique the required signal-to-noise ratio *(SNR_{req})* after demodulation to ensure an acceptable bit error rate (BER) will depend on the efficiency of any error correction coding scheme used. With a ‘desired’ signal of power *S* and uncorrelated (but not orthogonal) interference of power *I* at the receiver input, the signal-to-noise ratio after demodulation is

$$SNR = \frac{S}{I}G_p \tag{5.1}$$

or in dB

$$\begin{aligned} SNR\ (\text{dB}) &= S\ (\text{dB}) - I\ (\text{dB}) \\ &= G_p\ (\text{dB}) \end{aligned} \tag{5.2}$$

The amount by which *I* can exceed *S* is called the jamming margin, and is given by

$$\begin{aligned} JM\ (\text{dB}) &= I\ (\text{dB}) - S\ (\text{dB}) \\ &= Gp\ (\text{dB}) - SNR_{req}\ (\text{dB}) \end{aligned} \tag{5.3}$$

This figure is typically large, and it is often said that spread-spectrum signals are ‘hidden in noise’. Let us assume that N_u users are present, each received with equal strength *S*. Thus, in the absence of any other types of interference the total amount of interference seen from a given user is $I = (N_u - 1)S$. From (5.3), the maximum number of users will then be

$$N_u = JM + 1 \tag{5.4}$$

It is important to note that no other sources of interference have been considered here. If noise from other sources is dominant, or if no diversity schemes are being used in the presence of multipath interference, the capacity will be reduced from this figure.

5.2 Diversity schemes

A number of diversity schemes may be employed in DS CDMA, including multipath diversity and interferer diversity.

Multipath diversity refers to the ability to exploit self-interference. Multipath diversity may be employed due to the high bandwidth of the spread signal (in the range 1–10 MHz may be expected if DS CDMA is applied to UMTS). The result, as outlined in section 3, is that multipath reflections within a limited range may be exploited to improve the overall system performance in terms of bit error rate versus signal to interference ratio. In the case of DS CDMA this may readily be translated into increased system capacity, since many users simultaneously use the same frequency band. The direct translation between capacity and performance is intuitively an attractive feature of DS CDMA. For a TDMA scheme similar performance improvements may be achieved using equalisers. However, the performance gain may only be converted to capacity by decreasing the frequency reuse distance and hopefully the cell cluster size. This translation between performance and capacity is not necessarily practical in all cases for a TDMA scheme, thus giving DS CDMA an advantage.

Interferer diversity refers to the ability to average the effect of interfering signals on a large number of users. This is due to the correlation operation within the receivers, which correlates the desired signal and decorrelates interference. Interferer diversity when a large number of users are using the same frequency band is a major advantage of DS CDMA compared to TDMA. Whilst this is inherent in spread spectrum systems, carefully planned frequency hopping patterns are necessary in TDMA based schemes. No such planning is required in DS CDMA.

5.3 Adaptive power control

Adaptive power control on the reverse link (mobile to base station) is the single most important system requirement in DS CDMA applied to cellular systems [14]. The purpose of adaptive power control is to limit interference and transmitted power from the mobile station.

The importance of such a feature is easily recognised by considering the fact that in a DS CDMA system a large number of users may be sharing the same radio resources simultaneously. A signal received from a transmitter close to the receiver will tend to mask the signal from a more distant transmitter, if all transmitters transmit at the same output power level. This is known as the near-far problem. To cope with this, adaptive power control is used on the reverse link, with the aim of receiving all users at the same strength in the base station.

The speed, accuracy and dynamic range of the power control scheme are of prime importance to maintain capacity and radio link performance at acceptable levels. In general, the use of adaptive power control aims at tracking signal variations due to shadowing effects. A number of studies on the effects of imperfect power control has been carried out, e.g. [16]. The general conclusion to be drawn from these studies is that even relatively small variations around the perfect level (in the order of a few dBs) result in a significant decrease in overall system performance or capacity on the reverse link. Adaptive power control on the forward link does not have the same significance as adaptive power control on the reverse link. In practice, only a dynamic range in the order of a few dBs are realisable. This is due to the fact that reduced transmitted power on one code from a base station may significantly increase the interference level in mobile stations at cell boundaries. The strict requirements on power control on the reverse link is one of the major drawbacks of DS CDMA. Although the required power control performance may be achieved, this adds complexity to the system.

5.4 Bandwidth considerations

The 1992 World Administrative Radio Conference (WARC 92) identified 230 MHz of spectrum for use by future mobile systems (refer to UMTS article). However, not all of this spectrum will necessarily be available throughout the world for such use.

Given the need to support perhaps a large number of operators in the same geographical area, the need for full duplex transmission, the need for umbrella cells and the limited bandwidth available, it seems like the bandwidth requirement is a real constraint of DS CDMA. Obviously, other multiple access techniques require bandwidth in order to support a large number of users, but it is quite feasible to design narrowband FDMA or TDMA, whereas narrowband DS CDMA is a self-contradiction.

5.5 Complexity

Compared to more narrowband solutions, the bandwidth required to support DS CDMA systems implies an increase in receiver complexity, due to higher sampling rates and more stringent signal processing requirements. This translates into battery power requirements, which is a real constraint in a system where the majority of mobile stations are expected to be hand-held terminals.

The resulting complexity of the base station RF front end and decoding requirements due to the wideband nature of the DS CDMA signal should also be recognised. This applies in particular to multiuser receivers approaching performance levels anything near that of the optimal receiver. Moreover, the need for stringent power control schemes in order to maintain the desired capacity and performance of DS CDMA adds to the complexity of the mobile stations and base stations.

5.6 Radio sub-system planning and evolution

Frequency planning in a conventional TDMA or FDMA cellular system is costly and somewhat cumbersome. This could be avoided in such systems by introducing some form of dynamic channel allocation scheme, at the cost of major investments in the base station transceivers.

DS CDMA, on the other hand, allows cluster size of almost one, implying that almost all channels may be reused in every cell. Moreover, introducing new cells do not impact the frequency plan for a given area. TDMA based schemes rely on dynamic channel allocation strategies in order to avoid detailed frequency planning.

Concerning traffic planning, soft degradation of capacity is inherent in DS CDMA (refer to section 5.2), whereas performance increase and capacity do not directly translate in other multiple access schemes. Obviously, capacity may only be traded for performance up to a certain limit, as determined by the minimum performance requirements of the system. Once the overall performance degrades below a threshold, other strategies must be employed.

However, no matter what multiple access scheme is chosen, the use of umbrella cells calls for some degree of planning, as do coordination between multiple operators operating in the same geographical area or using adjacent bands.

5.7 Service and system evolution

DS CDMA is well suited for supporting a mixture of services at different and variable bit-rates. This includes services and transmission features such as packet data, variable bit-rate speech, voice activity detection and discontinuous transmission, etc.

The flexibility in introducing new services is very important, since it is not currently clear what services users will be requesting throughout the lifetime of a system. Thus, the inherent flexibility of DS CDMA allows scope for service evolution. The requirement to support evolution of transmission features and services also applies to other multiple access schemes, and is indeed possible. However, in DS CDMA the translation to available capacity is direct, whereas this is somewhat more cumbersome in the case of TDMA, as indicated above. In that sense DS CDMA inherently has a substantial potential for system evolution.

6 Summary

In this paper we have attempted to give a brief introduction to direct sequence spread spectrum techniques. The paper has focused on some of the main technical characteristics, capabilities and limitations of such techniques when applied to mobile radio systems. Following a description of the underlying principles of direct sequence multiple access some attention has been paid to some detector strategies appropriate in a mobile radio system, high-lighting some of the main characteristics of such detectors. Finally, we have briefly summarised some of the main features and potential consequences of applying direct sequence multiple access to mobile radio.

7 References

1 Proakis, J G. *Digital communications.* New York, McGraw-Hill, 1985. Series in Electrical Engineering.

2 Poor, H V, Verdu, S. Single user detectors for multi-user channels. *IEEE trans. comm.,* 36, 50–60, 1988.

3 Verdu, S. Minimum probability of error for asynchronous Gaussian multiple access channels. *IEEE trans. inform. theory,* 32, 85–96, 1988.

4 Lupas, R, Verdu, S. Linear multi-user detectors for synchronous code-division multiple access channels. *IEEE trans. inform. theory,* 35, 123–136, 1989.

5 Lupas, R, Verdu, S. Near-far resistance of multi-user detectors in asynchronous channels. *IEEE trans. comm.,* 38, 496–508, 1990.

6 Varanasi, M K, Aazhang, B. Multistage detection in asynchronous code-division : multiple access communications. *IEEE trans. comm.,* 38, 509–519, 1990.

7 Turin, G L. Introduction to spread-spectrum antimultipath techniques and their application to urban digital radio. *Proceedings of the IEEE,* 68, 328–353, 1980.

8 Price, R, Green, P E. A communication technique for multipath channels. *Proceedings of the IRE,* 46, 555–570, 1958.

9 Kaufmann, H, Kung, R. Digital spread-spectrum multipath-diversity receiver for indoor communications. In: *42nd vehicular technology society conference,* Denver, Colorado, IEEE, 1038-1041, 1990. ISBN 0-7803-0673-2/92.

10 Grob, U et al. Microcellular direct-sequence spread-spectrum radio system using N-path RAKE receiver. *IEEE journal on selected areas in communication,* 8, 772–780, 1990.

11 TIA/EIA Interim Standard. *Mobile station : base station compatibility standard for dual mode wideband spread spectrum cellular system.* Washington DC, 1993. (TIA/EIA/IS-95.)

12 Baier, A, Koch, W. Potential of CDMA for 3rd generation mobile radio systems. In: *MRC, Mobile Radio Conference,* Nice, France, Nov 1991, 13–15.

13 Lee, W C Y. Overview of cellular CDMA. *IEEE transactions on vehicular technology,* 40, 291–302, 1991.

14 Gilhousen, K S et al. On the capacity of a cellular CDMA system. *IEEE transactions on vehicular technology,* 40, 303–312, 1991.

15 Gilhousen, K S et al. Increased capacity using CDMA for mobile satellite communication. *IEEE transactions on selected areas in communication,* 8, 503–514, 1990.

16 Milstein, L B, Rappaport, T S, Barghouti, R. Performance evaluation of cellular CDMA. *IEEE transactions on selected areas in communication,* 10, 680–689, 1992.

Multipath propagation and its influence on digital mobile communication systems

BY RUNE HARALD RÆKKEN AND GEIR LØVNES

1 Introduction

In mobile communication systems, the quality of the radio link is the limiting factor of the quality of the end-to-end connection. To obtain a satisfactory quality of the radio link, it is of vital importance to have a good knowledge of the propagation conditions of the different environments where the systems will operate.

The radio energy propagates both directly from transmitter to receiver and via reflecting objects like hills, mountains, houses, etc. This mechanism is called multipath propagation. One important difference between analogue and digital communication systems is the consequence of this multipath propagation. While for an analogue system it does not matter how delayed the reflected signals are compared to the first arriving, a digital system may break down because reflected symbols reach the receiver within the next direct arriving symbol. In other words, multipath propagation may create intersymbol interference in a digital radio system.

We will start by giving a description of the mobile radio propagation channel. Next, the measurement equipment to monitor the propagation channel – a channel sounder – is described. Finally, results from different channel sounding measurement campaigns performed at 900 and 1700 MHz in macrocellular environments and at 1950 MHz and 59 GHz in microcellular environments will be presented.

2 The mobile radio propagation channel

2.1 Introduction

In mobile radio systems, all transmitter-receiver pairs transfer information from the input of the transmitter to the output of the receiver. The input of the transmitter can be a microphone or any data source, while the output of the receiver can be a loudspeaker or any data sink. Hence the transmitter-receiver pair represents a channel between the microphone/data source and the loudspeaker/data sink. In a mobile radio system the transmitter or receiver or both may be in motion during the information transfer.

In radio systems, the transmission medium is the electromagnetic spectrum. To transfer information, the transmitter will have to choose a carrier frequency from the spectrum, and then vary the amplitude, phase or frequency of this carrier frequency in sympathy with the signal from the information source. This process is known as modulation. At the receiver the information is extracted from the modulated carrier frequency in a demodulation process.

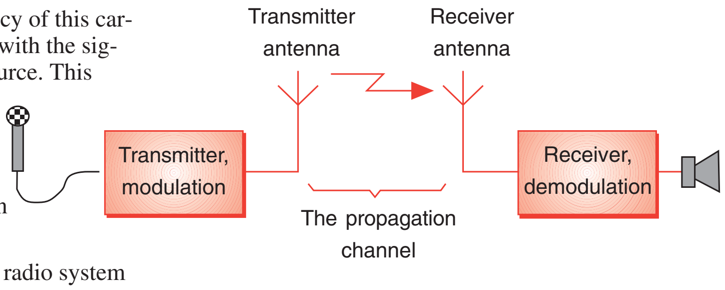


Figure 1 A radio system

In Figure 1 an example of a radio system is shown. The part of the radio system that is between the input of the transmitter antenna and the output of the receiver antenna, is called the propagation channel. In a mobile radio system this channel is constantly changing. The characteristics of the propagation channel depends upon the carrier frequency and on the amount of spectrum that is needed to transfer the information – the bandwidth of the propagation channel. Because energy will propagate from the transmitter to the receiver via reflecting objects in the environment, it is also strongly dependent on the environment in which the transmitter-receiver pair is situated. Finally, the propagation channel is influenced by the relative speed between transmitter, receiver and surroundings.

In this chapter we will assume that the transmitter is fixed, while the receiver is moving. We first look at the propagation channel as a function of position of the receiver, and start by transmitting only unmodulated carrier (carrier wave – CW). Next, we see what happens when the bandwidth is increased. Finally, we will introduce time, i.e. we let the receiver be moving with a certain speed.

2.2 The radio channel as a function of position

In Figure 2 we see a typical example of the situation of a radio connection. Both a direct signal and several reflected and diffracted ones are received – we have the situation of multipath propagation. Since the different signals have travelled different distances, the signals will be of different phase, amplitude and arrival time when they reach the receiver.

The received signal power of each path is approximately proportional to r^{-2}, where r is the distance between transmitter and receiver. But for the reflected and diffracted ones we will have additional attenuation in the reflection and diffraction points. This attenuation depends on the electrical properties (permittivity, permeability and conductivity) and surface roughness of the reflection materials.[1]

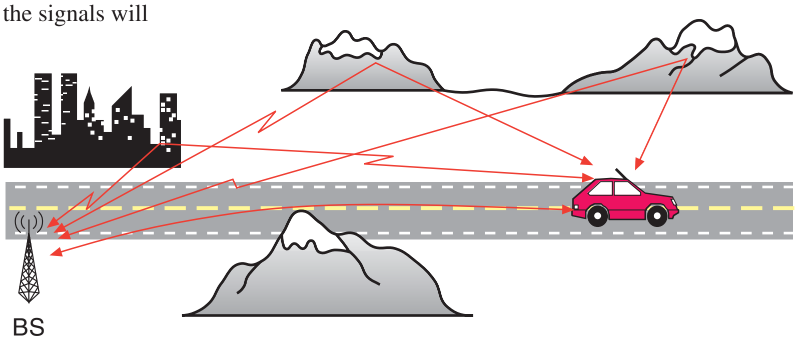


Figure 2 The situation of multipath propagation

In the coverage area the field strength will be very variable depending on the amplitude of the multipath signals and the degree to which the signals are in phase or out of phase. The reflection objects hence produce a standing wave pattern in the coverage area. This pattern consists of fades or holes of varying

[1] *Often in mobile communications it is assumed that the received signal power is proportional to $d^{-3.5}$ or d^{-4} where d is the distance from transmitter to receiver. This is however the total received signal, consisting of all rays, typically in situations where the direct path between transmitter and receiver is obstructed. Hence, all reflection and diffraction losses are included.*

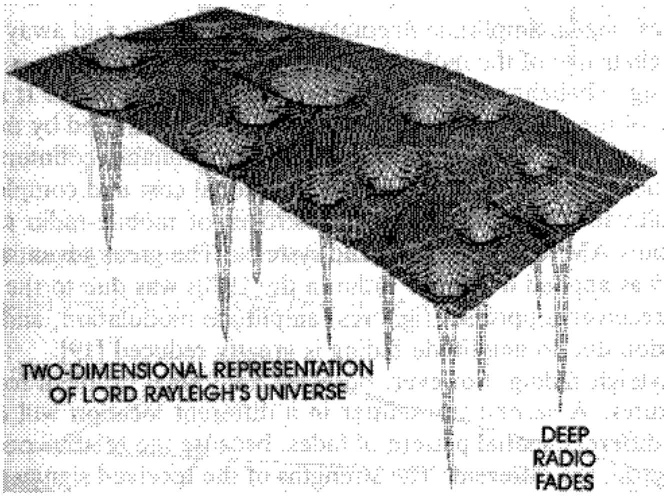


Figure 3 Reflection objects produce a standing wave pattern creating fades of varying depth [1]

depth – as seen in Figure 3. Typically, the fades may occur every λ/2 in any direction.

Moving in any direction in the standing wave pattern of Figure 3, or just jumping randomly from one position to another, we will see that the received signal amplitude approaches some distribution. If we receive equal amount of energy from all directions in the horizontal plane, i.e. all reflected signal components are equally strong, and no direct component is present, the amplitude becomes Rayleigh distributed. This can be shown mathematically by regarding the N incoming waves as independent random variables having the same distribution function with equal mean and variance. According to the central limit theorem the sum of the N stochastic variables approaches the Gaussian distribution for large values of N, independent of the individual distributions. Thus, the received signal can be modelled as an in-phase and quadrature component Gaussian random process. If x_1 and x_2 are the two Gaussian processes,

then $\quad r = \sqrt{x_1^2 + x_2^2}$

is the envelope and it can be shown that r is Rayleigh distributed with mean

$$\sqrt{\frac{\pi}{2}} \cdot \sigma \text{ (= constant).}$$

If the direct component is present the Rice distribution has been proposed. (The Rayleigh distribution is a special case of the Rice distribution.)

In reality the Gaussian components x_1 and x_2 will have a slowly varying mean value due to terrain variations. Then our amplitude distribution model is a Rayleigh (or Rice) distribution with a varying mean. It seems that a log-normally distributed mean compare well with measurements.

The standing wave pattern of Figure 3 is specific for one frequency – we assume CW has been transmitted. If the transmitting frequency is changed we will have another pattern with fading dips in other positions. This is because the differences between the multipaths have changed in terms of wavelengths, so the spots where the multipath signals cancel each other out are not the same as before.

When all reflection objects are close to the direct path between transmitter and receiver, the difference in wavelengths between the shortest and longest path will be small. Then we must make a major change of the transmitter frequency to get large variations in the standing wave pattern. On the other hand, when this difference is large, we do not need to change the frequency much to get large variations in the standing wave pattern.

By measuring the correlation between the two signals (frequencies) we get a very important parameter of the mobile radio channel: The coherence bandwidth, B_c. It is often defined as the necessary frequency separation to make the correlation between the two signals equal to 0.5.

Coherence bandwidth is illustrated in Figure 4, and it is an important parameter because it determines whether a system is narrowband or wideband. If the bandwidth of a system is greater than the coherence bandwidth, the system is exposed to frequency selective fading, and the system is called a wideband system. On the other hand: If we do not have frequency selective fading in our system, then the bandwidth is less than the coherence bandwidth, and we have a narrowband system. This situation with no frequency selective fading is called a flat fading situation.

In a narrowband system we will find Rayleigh fading with deep fades. In a wideband system each frequency component exhibits this Rayleigh fading nature, but the probability that all frequency components will fall into a deep fade simultaneously is very small. Thus in a wideband system we seldom experience deep fades; generally, the amplitude variations of the wideband system is much smaller than in the narrowband system. When the bandwidth of a system is increased, the overall fading behaviour thus asymptotically approaches the log-normal situation.

Another important parameter, which is closely related to the coherence bandwidth, is the delay spread. To find the delay spread, rather than transmitting CW, we must perform wideband measurements to record the impulse response of the channel. This may be done by transmitting a short pulse. The pulse will propagate via the reflecting objects, and at the receiver we will receive many pulses – echoes which are delayed and attenuated. If we plot the power of each received echo with excess delay relative to the first received signal, we have found the power delay profile or impulse response of the channel. The parameter delay spread is defined as the power weighted standard deviation of the impulse response (Figure 5). The effect of delay spread is to smear or spread out the signal. If the symbol duration is shorter than the delay spread, we get

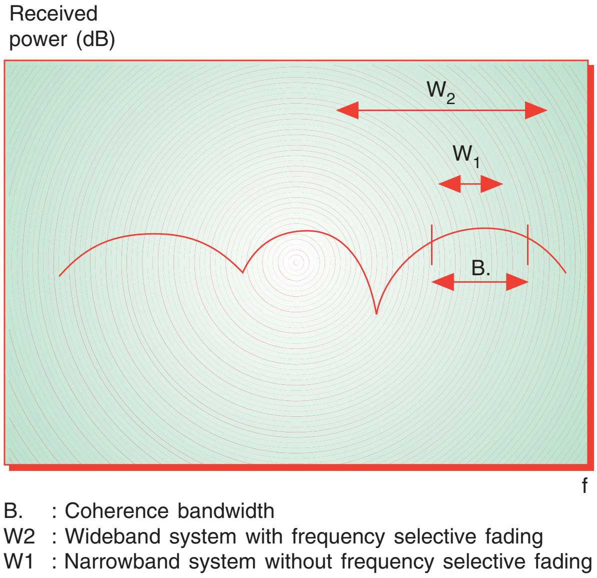

Received
power (dB)
W2
W1
B.
f
B. : Coherence bandwidth
W2 : Wideband system with frequency selective fading
W1 : Narrowband system without frequency selective fading

Figure 4 Illustration of coherence bandwidth

intersymbol interference – the symbols overlap.

The delay spread is inversely proportional to the coherence bandwidth. This roughly implies that if the symbol duration is shorter than the delay spread, we have both intersymbol interference and frequency selective fading. And the opposite: When the symbol duration is longer than the delay spread, we have flat fading and no intersymbol interference.

So far, we have discussed the amplitude distribution of the received signal. What about the phase and frequency?

Regarding phase, there is no reason for one particular phase to appear more often than another. Thus the phase probability function is equally distributed within the interval 0 to 2π. About the frequency distribution we can say nothing at this point, because so far we have made no assumptions about time. Since frequency is the time derivative of phase, we know nothing about it unless time and the velocity of the receiver is introduced. This is what we are going to do next. This also enables us to present the power spectra of the received signal.

2.3 The radio channel as a function of time

In the previous section we considered the observed signal in the coverage area of a transmitter without assuming anything about how we moved around. At each location we received a huge number of waves of equal frequency but with different amplitudes and phases. If we now instead put our receiver in a moving vehicle, we will see the same huge number of incoming waves, but each with a Doppler shifted frequency. The Doppler shift of each ray corresponds to the suppression or stretching of the radio waves when we are driving towards or away from the ray source. We will find the fades at exactly the same positions as before and the distribution of the amplitude remains unchanged. The Doppler shifts of the received rays, however, will be experienced by the receiver as random frequency modulation.

Generally, if α_n is the angle between the direction of the movement and the n-th incoming wave, the Doppler shift of this component is:

$$f_n = \frac{\nu}{\lambda} \cdot \cos\alpha_n \tag{1}$$

The maximum Doppler shift is

$$f_n = \frac{\nu}{\lambda} \tag{2}$$

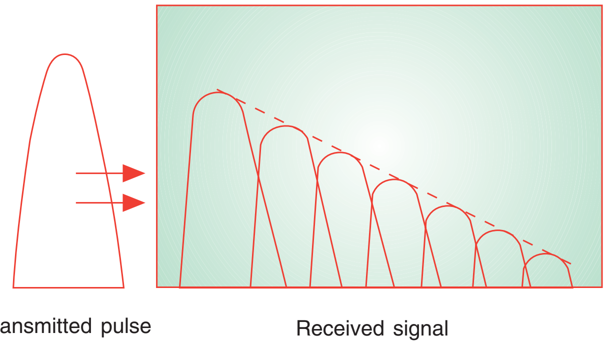


Figure 5 The delay spread is commonly used to describe the difference in time between the first and last received pulse. The mathematical definition of the delay spread parameter is the standard deviation of the envelope of the impulse response

The inclusion of time and movement also enables us to detect power spectra.

If CW is transmitted, the received signal can be expressed as:

$$r(t) = a(t) \cdot e^{j\Psi(t)} \tag{3}$$

where $a(t)$ is the envelope of the signal due to the fading and $\psi(t)$ is the phase term. $a(t)$ can be separated into two terms: $a(t) = a_l(t) \cdot a_r(t)$ the log-normal- and Rayleigh/Rice fading factors respectively. Thus, when extracting the long-term fading factor, we have:

$$r_r(t) = a_r(t) \cdot e^{j\Psi(t)} \tag{4}$$

The power spectrum of $r_r(t)$ is called the RF Doppler spectrum or the signal spectrum. This is the spectrum at the input of the receiver (i.e. at the output of the receiver antenna.). If uncorrelated wave components of equal mean power are received from all directions in the horizontal plane, and the receiver antenna is a vertical monopole, we will get an RF Doppler spectrum on the receiver's input as shown in Figure 6a. If one dominant component is arriving at an angle α_0, the RF spectrum will be as shown in Figure 6b.

The corresponding baseband power spectra – the spectra of $a_r(t)$ – are shown in Figure 7. Note that these spectra drop to zero at twice the maximum Doppler frequency.

We now move on to the $e^{j\Psi(t)}$ term, the phase term. The random frequency modulation described earlier is the time derivative of this phase,

$$\Psi(t) = \frac{d\Psi(t)}{dt}.$$

In Section 2.2 we stated that the phase was uniformly distributed from 0 to 2π, but we were not able to say anything about the speed of the phase shifts. Rapid changes are associated with deep fades of the signal. With infinitely deep fades, we can have infinitely fast phase changes. Hence there is always a possibility of

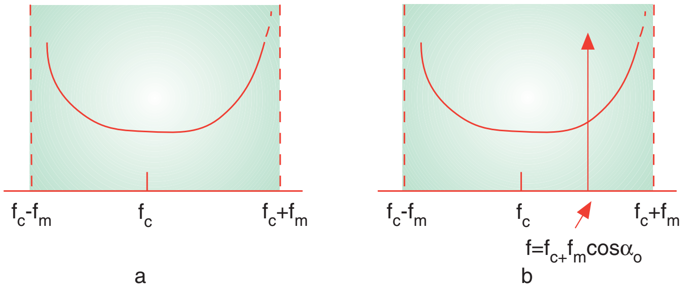


Figure 6 RF power spectrum of the received signal, (a) with equal mean power received from all directions, (b) with an additional dominant component received from an angle α_0

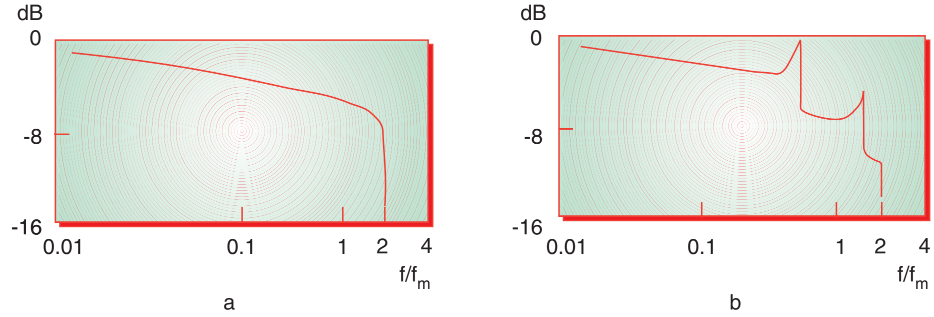


Figure 7 Power spectrum of the signal envelope. (a) and (b) correspond to the situations in Figure 6

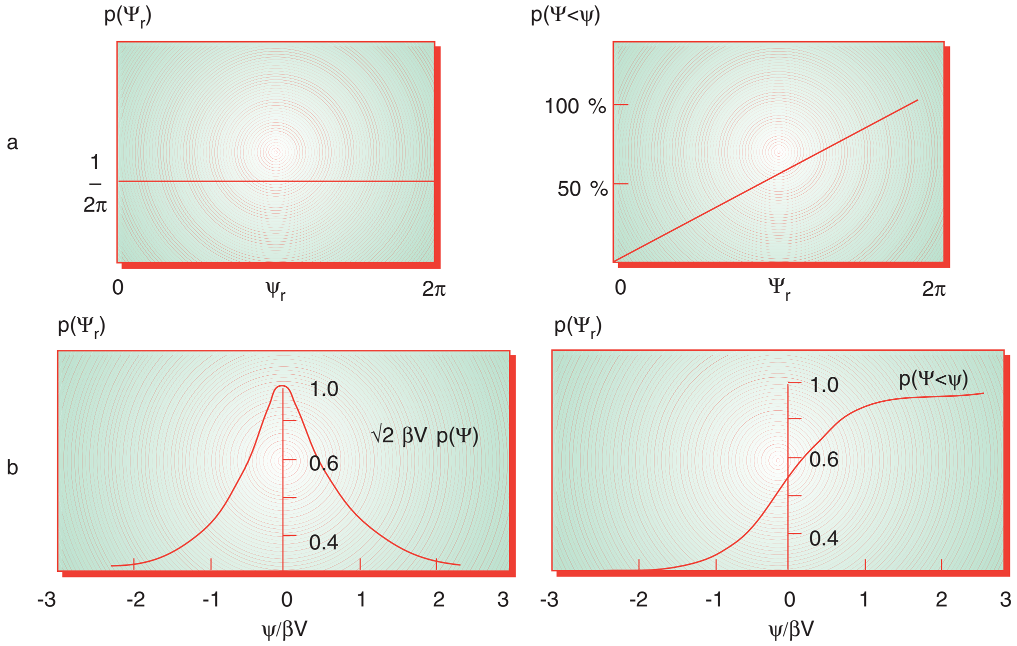


Figure 8 Probability densities and cumulative distributions for the random phase (a) and random frequency (b) modulation

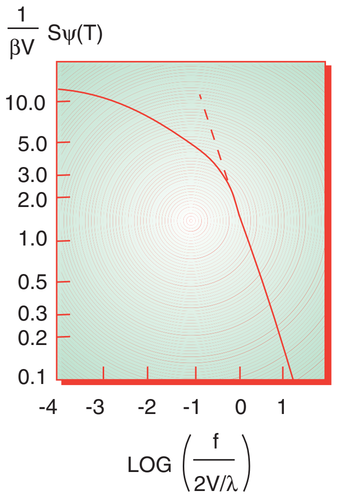


Figure 9 Power spectrum of random FM

finding any frequency of the random FM, but the most probable is the carrier frequency. The probability densities and cumulative distributions for the random phase and the random frequency modulation of the signal are shown in Figure 8.

In Figure 9 the power spectrum of the random FM is shown. It is a function of the maximum Doppler frequency. We see that the spectrum never reaches zero. Above twice the Doppler frequency the spectrum falls off as 1/*f*. This frequency is thus regarded as a cut-off-frequency for random FM, and it is identical to the cut-off-frequency for the baseband power spectrum.

The derivation of these spectra can be found in the literature, for instance [2] or [3].

In section 2.2 we looked at the mobile propagation channel when both receiver and transmitter were stationary, and we started by transmitting CW. Next we instead transmitted a short pulse to see the time dispersion of the static channel. Then we went back to CW, allowing the receiver to be in motion. Due to the reciprocity principle it does not matter whether it is the receiver or transmitter that is in motion. To complete our description of the mobile propagation channel, we thus transmit a short pulse and let the receiver move.

What then happens is that the received power delay profile becomes a power delay-Doppler profile, i.e. every received ray is characterised not only by its power and excess delay, but also by its Doppler shift. In real life we are not able to transmit an infinitely narrow pulse, hence every received pulse represents the sum of all rays that have been reflected from points within ellipsoid shells with the receiver and transmitter as foci (Figure 10). The thickness of the shells corresponds to the width of the transmitted pulse, and rays reflected from one shell will fall into the same time bin of the power delay profile. If a large number of reflection sources is found within a shell, the received signal amplitude of this bin may be Rayleigh (or Rice) distributed, and the Doppler spectrum of this bin may look like the spectra in Figure 6.

When wideband measurements are performed, the delay-Doppler profiles can be recorded. If the transmitter and receiver are not frequency synchronised (e.g. by reference oscillators) only power delay profiles can be obtained. A huge amount of profiles can be recorded along a measurement route, and often some reduction of data must be done. One way to do this is to make typical delay-Doppler profiles based on the recordings. These are characterised by the number of bins and the delay, average power, fading characteristics and Doppler spectrum of each bin. These profiles will reflect the situation in specific environments. Another method is to calculate important statistical parameters based on the power delay profiles and show cumulative distributions of these parameters. One such parameter is the delay spread which we have already introduced. In Section 3.2.4 other statistical parameters are presented.

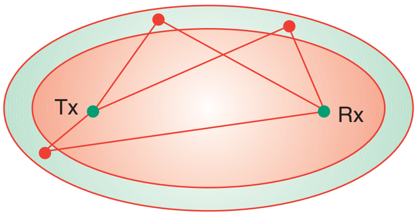


Figure 10 All rays that have been reflected from points within one ellipsoid shell corresponding to a certain time bin of the impulse response will contribute to the received power of this time bin

3 Channel sounding techniques

There are two classes of equipment for measuring propagation conditions of the mobile radio channel. Narrowband equipment transmits and receives CW (or a frequency spectrum much smaller than the coherence bandwidth). This class is mainly used for path loss measurements. Wideband equipment – channel sounders – transmits a frequency spectrum above the coherence bandwidth, and is used for channel characterisation.

3.1 Different types of channel sounders, benefits and limitations

Channel sounders are often categorised into three main classes: pulse, pseudo-random sequence and frequency sweep (chirp) sounders. The two latter use pulse compression techniques.

In a pulse channel sounder, a short RF-pulse is transmitted, and the received signal envelope is detected in the receiver. Only information about the received signal amplitude is obtained, it is thus not possible to get any information about the Doppler spectra. Unless very large transmitter power is used, the distance between transmitter and receiver is limited. In addition this method is sensitive to interference from other services.

The signal used for sounding the channel by a pseudo-random sequence sounder, is the carrier modulated with a pseudo-random binary sequence. In the receiver a sliding correlator, a signal processor performing correlation or a channel matched filter is used to estimate the channel impulse response. Using this technique it is possible to obtain a larger range between transmitter and receiver. If Doppler spectra are to be derived, frequency synchronisation between transmitter and receiver is needed.

The frequency sweep technique is well known from radar theory. Traditionally, in the receiver a pulse compression filter is matched to the transmitted wave form. The resolution of the measurement system is inversely proportional to the bandwidth of the frequency sweep, and the maximum measurable delay is proportional to the duration of the sweep. Also this measurement technique makes it possible to obtain large ranges between transmitter and receiver even when moderate output power is used. In addition, the method is resistant to interference from other services and gives good utilisation of the used measurement bandwidth.

3.2 Telenor R&D's channel sounders

Telenor Research and Development has performed wideband measurements on UHF and at 59 GHz. Two different channel sounders were used, and RF-parts were changed in accordance with the desired radio frequency. Both channel sounders are based on the frequency sweep technique. The channel sounder used for measurements in the millimetre wave band (from now on mentioned as the 200 MHz channel sounder[2] is an evolution from the old 8 MHz channel sounder.

The signal source is a set of two digital frequency sweep generators which provide the in-phase and quadrature signal components. The signal is converted to analogue and mixed up to an intermediate frequency before entering the RF mixer and amplifiers.

The output signal is shown in Figure 11. The transmitted wave form is a frequency sweep (chirp) from $f_0 - \Delta f$ to $f_0 + \Delta f$ (1 MHz ≤ $2\Delta f$ ≤ 8 MHz for the 8 MHz channel sounder, 6.25 MHz ≤ $2\Delta f$ ≤ 200 MHz for the 200 MHz channel sounder), with a repetition frequency inversely proportional to the sweep duration.

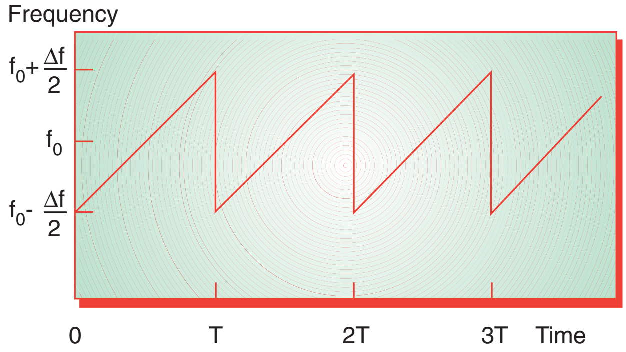


Figure 11 The transmitted signal

Transmitting the chirps successively makes it possible to use a class C power amplifier.

In the receiver the signal is down-converted to baseband and sampled. The samples are stored in RAM and transferred to a digital signal processor (DSP) performing a Fast Fourier Transform (FFT) of the received chirp. This signal is then correlated with a Fourier transformed replica of the transmitted chirp (multiplication in the frequency domain), previously weighted by a weighting function to reduce sidelobes in the resulting channel estimate. Finally, an inverse FFT is carried out and the result is an estimate of the radio channel's complex impulse response.

In both receivers several sweeps can be added (up to 31 in the 8 MHz sounder and up to 682 in the 200 MHz channel sounder) before channel estimation takes place. This averaging lowers the receiver noise floor significantly. For instance, if ten succeeding chirps are added before channel estimation takes place, the signal to noise ratio is improved by 10 dB.

An overview of the channel sounders is given in Figure 12.

During the channel measurements, in-phase and quadrature (I and Q) impulse response samples or delay-Doppler spectra (200 MHz sounder only) are transferred to the PC hard disc together with received power and system information. An oscilloscope is used as an on-line system status or impulse response monitor, allowing real time channel observations. A Multitrip is used to register informa-

[2] *To be able to distinguish between the two different channel sounders owned by Telenor R&D, the channel sounders are from now on mentioned as the 200 MHz channel sounder and 8 MHz channel sounder. The naming convention is reflecting the maximum measurement bandwidths of the two channel sounders.*

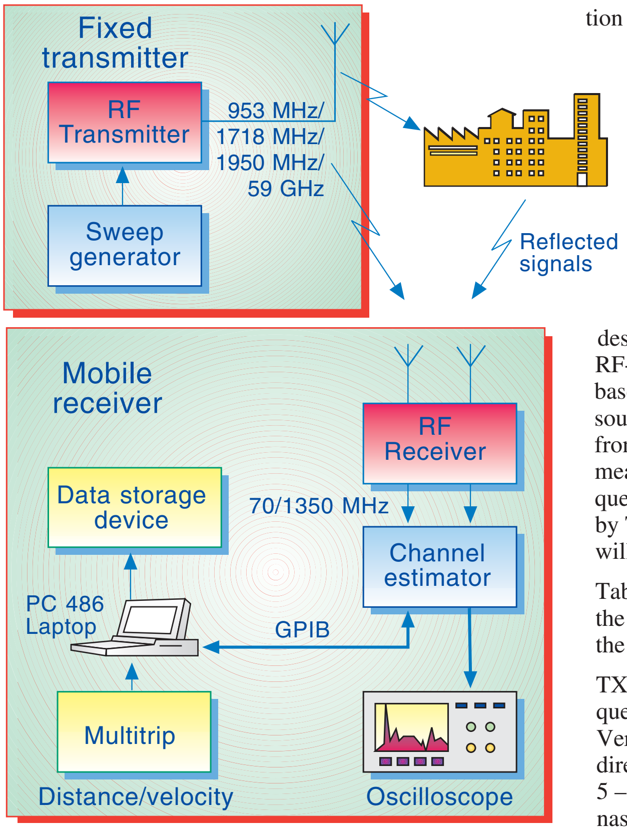


Figure 12 Overview of the channel sounders

Table 1 8 MHz channel sounder parameters

Sweep length, µs	128	64	32	16
Sweep BW, MHz	1.0	2.0	4.0	8.0
IR length, µs	100	50	25	12.5
IR resolution, µs	1	0.5	0.25	0.125

Table 2 Main parameters of the 200 MHz channel sounder

Parameter	Value/range
Bandwidth	6.25 – 200 MHz
Instantaneous dynamic range	> 30 dB
Total dynamic range	> 80 dB
Channel sampling rate	max 728 IRs/s
IR resolution	5 – 160 ns
No. of samples per IR	64 – 8192, complex
Chirp duration (IR length)	0.32 – 163.84 µs
Number of averaged chirps	1 – 682 chirps

tion about travelled distance and velocity when used in a vehicle.

The digital parts of both channel sounders are developed by the Norwegian research company SINTEF DELAB according to specifications given by Telenor R&D.

3.2.1 The 8 MHz channel sounder

The 8 MHz channel sounder is described in [5]. The mixers and the RF-parts are not integrated with the baseband and IF-parts of the channel sounder, making it easy to change RF front-ends to perform channel sounding measurements at different radio frequencies. RF-parts have been assembled by Telenor R&D. A brief description will be given in the following.

Table 1 shows how the parameters of the channel sounder vary depending on the sweep length:

TX power is 25 W for both carrier frequencies (953 and 1718 MHz) used. Vertically polarised omnidirectional and directional transmitting antennas with 5 – 15 dBi gain are used. The RX antennas have approximately 5 dBi gain.

Instantaneous and total dynamic range of the receiver is more than 30 and 80 dB respectively, and 15 IRs are estimated every second.

Sampled impulse responses (128 complex samples in each) are stored on the disk of a PC, and channel statistics are calculated afterwards.

The receiver sensitivity is defined as the input level where the instantaneous dynamic range of the receiver is 0 dB. Receiver sensitivities for the 8 MHz channel sounder vary between –130 dBm (8 MHz BW and no averaging) and –145 dBm (1 MHz BW and averaging of 31 received sweeps before channel estimation).

3.2.2 The 200 MHz channel sounder

The 8 MHz channel sounder has been redesigned into a new sounder designed to suit measurement demands in the millimetre wave band. This sounder is capable of performing impulse response measurements at a best resolution of 5 nanoseconds, corresponding to a measurement bandwidth of 200 MHz. The sounder also incorporates the possibility of narrowband path loss measurements, and signals from two separate receiver antennas may be sampled simultaneously to be able to evaluate the benefits of space diversity. The RF part is not integrated with the IF- and baseband parts of the channel sounder, hence it is possible to use different RF front-ends to perform channel sounding measurements at different frequency bands.

The 59 GHz RF part includes a transmitter with an output power of 500 mW, with a choice of using either omnidirectional or directional antennas. The former are bi-conical horns with vertical half power beam width of 20 degrees. The latter is a horn antenna with horizontal and vertical beam widths of 90 and 20 degrees respectively. The antennas are vertically polarised.

To achieve reliable results, the channel sounder is capable of equalising the transfer function of the entire measurement set-up.

Receiver sensitivities are variable from –110 dBm (200 MHz BW, no averaging) to –141 dBm (6.25 MHz BW, averaging of 42 received sweeps).

The 1950 MHz transmitter has an output power of maximally 10 W. Receiving antenna is an omnidirectional λ/4 dipole antenna with 3 dBd gain. Transmitting antenna is either identical to the receiving antenna, or a directional antenna with 13 dBd gain.

Due to 7 dB lower noise figure of the 1950 MHz RF front-end than the 59 GHz RF front-end, the receiver sensitivities are also 7 dBs better than at 59 GHz.

Main parameters of the channel sounder are given in Table 2.

MA/COM, USA, has provided the 59 GHz RF parts. The 1.95 GHz parts have been assembled by Telenor R&D. In addition, 5.2 GHz RF front ends are under assembly.

3.2.3 Propagation data analysis

Telenor R&D has developed a comprehensive data program to control the channel estimator and collect measurement data [9]. Analysis of the stored impulse responses is done off-line by this program system.

The program determines a noise spurious threshold (NST) based on the receiver sensitivity applying for the selected measurement set-up. The NST is used to determine whether the components in the estimated impulse response are real signal reflection components or due to thermal or correlation noise. Impulse responses disturbed by low signal to noise ratio or aliasing caused by too long multipath delays, are identified and will not be considered in the data analysis.

Important statistical parameters describing the impulse responses are calculated by the computer and may be displayed graphically as cumulative distributions. The most common parameters are [4]:

- ***Mean Delay (MD):*** The first order moment of the IR; the power-weighted average of excess delays.
- ***Delay Spread (DS):*** The second order moment of the IR; the power-weighted standard deviation of the excess delays.
- ***Fixed Delay Window (FDW):*** The length of the *middle* portion of the IR containing a certain percentage of the total energy of the IR.
- ***Sliding Delay Window (SDW):*** The length of the *shortest* portion of the IR containing a certain percentage of the total energy of the IR.
- ***Delay Interval (DI):*** The interval between the first time the power of the IR exceeds a given threshold and the last time it falls below the threshold.

The five parameters describe the statistical properties of *the channel impulse response* itself.

By averaging the IRs obtained over a distance corresponding to some tenths of wavelengths, we get the so-called (average) power delay profile. During the averaging process the fast fading of every component of the IR will be eliminated. Then we can calculate the averaged parameters using the power delay profile as weighting function rather than the IR.

It is often interesting to evaluate parameters that give some indication of how well a digital radio system will work in a given multipath environment, and how damaging multipath propagation is for the communication.

For a given modulation method it is possible to calculate the bit error rate as a function of delay spread, assuming that a channel equaliser is not present in the receiver. There is for instance a rule of thumb telling that in a system without channel equaliser, using GMSK-modulation, there will be an irreducible bit error rate (BER) of approximately 10^{-3} if the delay spread parameter exceeds one tenth of the symbol interval.

A BER of 10^{-3} is defined as performance limit in DECT (Digital European Cordless Telecommunications). DECT is one of the strong candidates to radio in the local loop (RLL), where the idea is to replace parts of the copper network with radio solutions. A DECT implementation not employing a channel equaliser is in accordance with the General Access Profile (GAP) in the DECT specifications. Hence the delay spread gives a direct measure of the performance of a DECT implemented according to the GAP. The bit interval in DECT is 868 ns.

Since situations with multipath components having excess delays of more than one symbol interval are common in many environments, some radio systems use receivers equipped with a channel equaliser to work properly. This is the case for GSM (and DCS 1800) having a symbol duration of 3.7 µs.

The GSM and DCS 1800 receivers are designed to cope with multipath profiles stretching over 4 symbol intervals, and the *Q_{16}-parameter* has been suggested to indicate how well these systems will work under given multipath conditions. The Q_{16}-parameter is defined as follows:

- ***Q_{16} ratio:*** The ratio of the power inside to the power outside a window of duration 16 µs. For each IR the window is slid to find the position with highest power inside the window.

The idea behind the Q_{16}-parameter is that the equaliser is useful in one out of three situations: If the multipath components are received within one data symbol duration, the receiver bandwidth is insufficient to resolve the components, and flat fading will occur (situation I). In this situation the equaliser is of no use. If significant components are received with excess delays larger than the equaliser depth, these signal components cannot be utilised by the channel equaliser, and they hence represent intersymbol interference. This may cause severe problems for the communication (situation III). But if the significant multipath components are received with excess delays less than the equaliser depth, but larger than the symbol duration, the equaliser will improve the performance by exploiting the energy in the multipath components (situation II).

We assume that GSM will work satisfactorily in a specific environment in situations I and II, but not in III. According to the GSM specifications, a GSM receiver will work properly if the average carrier

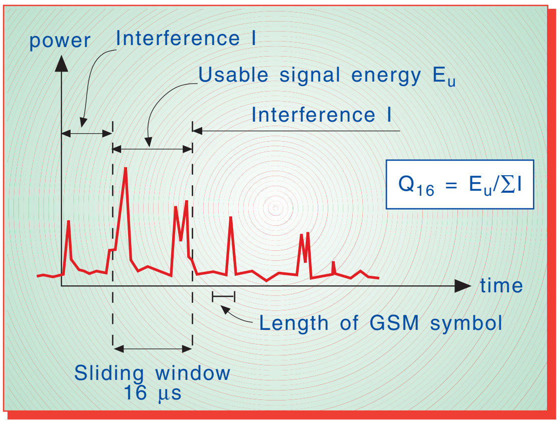


Figure 13 Illustration of the Q_{16}-parameter

to co-channel interference level is better than 9 dB. Co-channel interference may be intersymbol interference or interference from GSM transmitters in other cells. This means that if there is no interference from other cells, the received energy falling within the equaliser window must be at least 9 dB above the energy falling outside this window. In GSM the equaliser depth is almost 15 µs. (For historical reasons the parameter has been defined assuming an equaliser window of length 16 µs.)

From the description above, we see that the Q_{16} and sliding delay window parameters are based on the same ideas, giving an indication of the performance of a radio system equipped with a channel equaliser in a given multipath environment. The main difference is that Q_{16} is a parameter suited to describe performance of a radio system with known channel equalising capabilities. The SDW parameter, on the other hand, is suited to determine the channel equalising capabilities needed to make a radio system with given characteristics work in a given environment. Hence, we can say that Q_{16} is system dependent, whereas sliding delay window is system independent.

The reason for presenting both fixed and sliding delay window is that the fixed delay window (or just delay window) is a well-established parameter that is useful when comparing the results presented by different authors.

4 The mobile radio channel in different environments

The effort of the propagation measurements activities at Telenor R&D has until now been twofold:

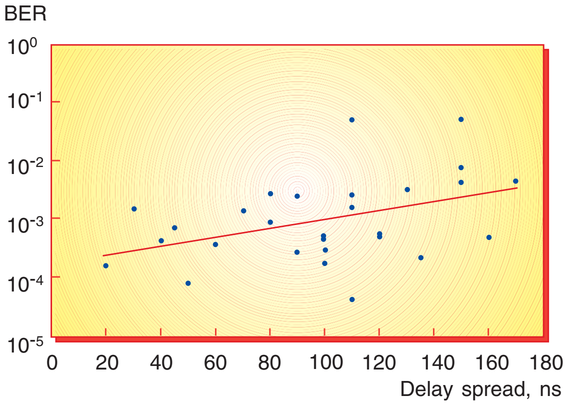


Figure 14 Bit error rate obtained in DECT versus averaged delay spread. Measurements with received power close to the sensitivity limit have been discarded

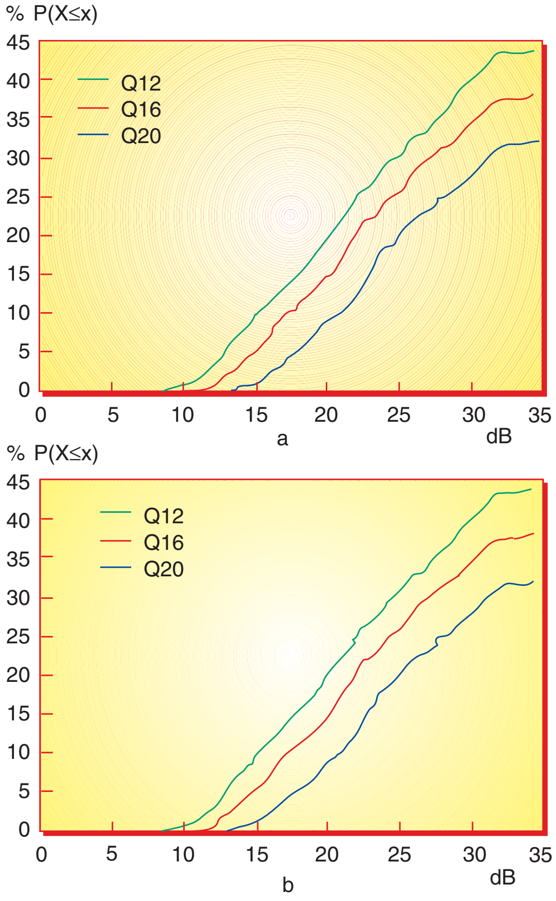


Figure 15 Cumulative distributions of average Q-parameters for Farm_1
a) 953 MHz
b) 1718 MHz

- Exploring propagation conditions at 900 and 1700 MHz in *macrocells*. In addition, we have performed parallel multipath- and GSM signal quality measurements to verify the suitability of the Q_{16}-parameter as GSM signal quality estimator
- Exploring propagation conditions at 1950 MHz and 59 GHz in *microcells*. We have also performed combined multipath- and DECT BER measurements to verify the suitability of the delay spread parameter as a BER estimator in a system not employing a channel equaliser.

We will present some results from measurements in different environments and within different frequency bands.

4.1 Multipath and GSM signal quality measurements

Parallel multipath and GSM measurements have been conducted to give an indication of how well suited the Q_{16}-parameter is for estimating GSM signal quality, and thereby also estimating GSM quality of service.

In given environments along 5 measurement routes, coverage prediction based only on received signal level indicates full coverage, whereas the GSM signal quality was unacceptable in up to 52 % of one specific measurement route. Also taking multipath effects into account when estimating GSM coverage by use of the Q_{16}-parameter with a fixed threshold of 9 dB gave a correct estimate in 82 % of the cases. Using a 3 dB separation between Q_{16} thresholds for acceptable and unacceptable signal quality, the GSM quality was estimated with an accuracy of more than 90 % (with 7 dB as lower threshold, 10 dB as upper) [11]. Hence, multipath measurements prove to be a useful tool for estimating GSM coverage in areas believed to have destructive multipath conditions.

4.2 Multipath and DECT BER measurements

Combined multipath and DECT measurements have also been conducted. Discarding the measurements where the received signal levels were close to the DECT sensitivity limit, we found that even though there was a limited set of measurements spread around the regression line, there was a strong correlation between the values of averaged delay spread and BER obtained in the DECT system [12]. Well in line with the rule of thumb, the DECT performance limit of BER 10^{-3} was reached for delay spread values around 100 nanoseconds, as shown in Figure 14.

4.3 Multipath measurements in macrocells at UHF frequencies

Different parameters based on both instantaneous and averaged impulse responses were calculated [6], and since those macrocell measurements mainly are aimed at GSM, we will focus on the Q_{16}-parameter (based on averaged impulse responses, because it is the average C/I_c that is to be 9 dB or more for GSM to work properly) and how multipath propagation will influence the transmission quality of a digital radio system like GSM. To give an indication of the influence of the equalising capabilities, we will also show corresponding figures for the parameters Q_{12} and Q_{20}.

Traditionally, coverage of cellular systems has been predicted taking only received signal power and interference level into account. In digital radio systems, effects of multipath propagation also needs to be taken into account to give reliable coverage predictions.

For some measurements, we also show examples of typical IRs. These examples are IRs that we feel are representative for the particular scenario, and we have chosen them after looking thoroughly through the recorded IR files. We never show IRs recorded at 953 MHz and 1718 MHz along the same measurement route, because we do not want the reader to do comparisons between the two frequencies based on single impulse responses only.

We want to stress that all typical IRs shown from macrocell measurements in the UHF band are under non line of sight (NLOS) conditions. With line of sight (LOS) we seldom saw more than one path – the reflected ones were typically more than 30 dB attenuated compared to the direct one.

For most of our macrocell measurements we chose already existing NMT base station sites as transmitter sites. In those cases, the height of our transmitting antenna was usually lower than the heights of existing NMT base station antennas.

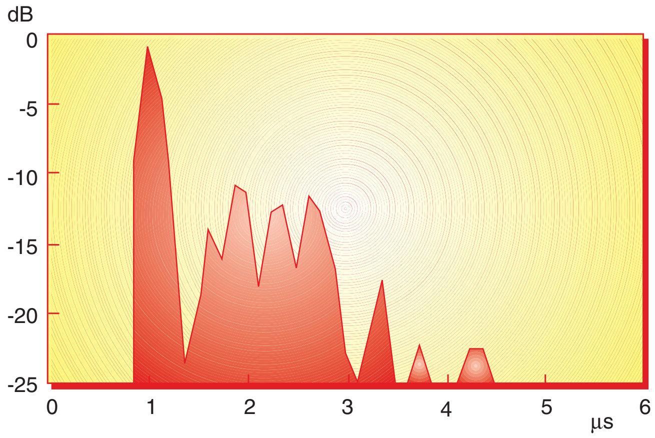


Figure 16 Typical IR from Farm_1, 1718 MHz. Total received power was –84 dBm, 20 dB above GSM sensitivity level

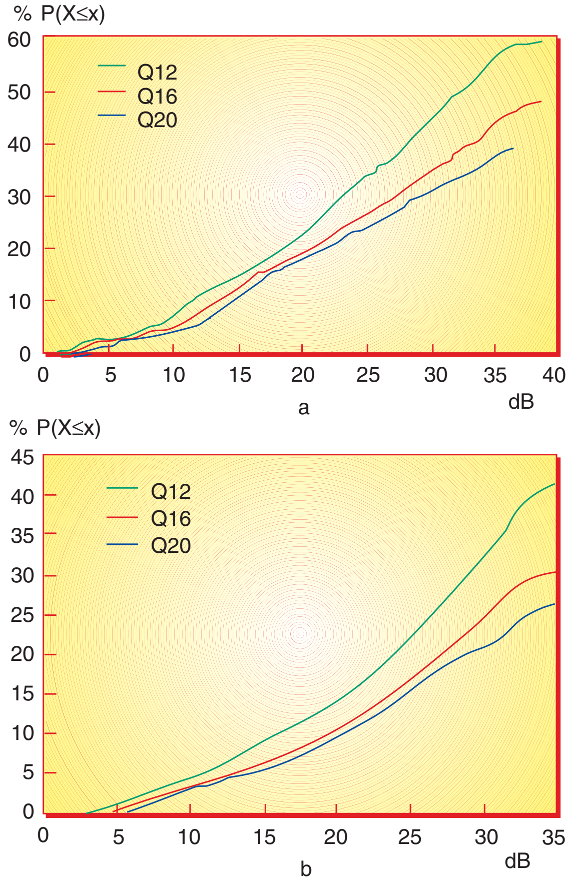


Figure 17 Cumulative distributions of averaged Q-parameters for Farm_2
a) 953 MHz
b) 1718 MHz

4.3.1 Measurements in rural farmland and forest

Measurements were conducted from 5 different transmitter sites in rural areas with varying degree of open farmland and forest. In general, for all measurements performed in this environment, typical lengths of IRs under NLOS conditions were 10 µs, sometimes increasing to 20 – 40 µs. From a total of 6 measurement routes, Q parameters from two routes are presented.

Farm_1: An open rural area with gently sloped farmland. Scattered residential houses and trees along the measurement route. Antenna height 8 m above ground level. The measurement route was 4.9 km long with a distance of 0.6 – 5 km from TX. The direct LOS path was often obstructed by the smoothly rolling terrain. During NLOS conditions, the longest IRs were found closest to the transmitter. From Figure 15 we see that the multipath situation is more severe at 1718 than at 953 MHz. Multipath propagation will cause no problems for GSM and DCS 1800 along this route.

Farm_2: The transmitter placed in an open, rural environment with farmland and some forests. The antenna was placed 8 m above ground level. The distance to the measurement route was 0.3 – 7 km. The route was 16 km long and went through farmland and forests. The forest and small, smooth hills caused an NLOS situation along the route. The reflections were generated by other wooded 100 – 150 m high hills in the area. Along this route several of the received multipath components were far below the sensitivity level of GSM. From Figure 17 we see that multipath propagation may cause problems for GSM and DCS 1800 in between 5 and 10 % of the route.

4.3.2 Measurements in valleys

Measurements have been performed from four different transmitter sites in different valleys, with a total of 5 measurement routes. Results from two routes are presented here.

Val_1: A rural area in a smooth “U”-shaped valley, 3 km wide with 350 m high wooded hillsides and a 2 km wide lake in the middle. The TX was placed in the hillside 230 m above the measurement route. The route followed the lake at a distance 1.3 – 12 km from the TX. The direct LOS path was generally obstructed by the uneven hillside. The length of the IRs were 20 – 40 µs, even right below the transmitter.

We see from Figure 19 that a GSM or DCS 1800 receiver is not able to communicate with a base station situated at the chosen transmitter site. Thus, if this were the only possible base station in the area, the systems would fail to work. It is worth mentioning that this base station site has been used for the analogue NMT system for years, always giving satisfactory quality of service.

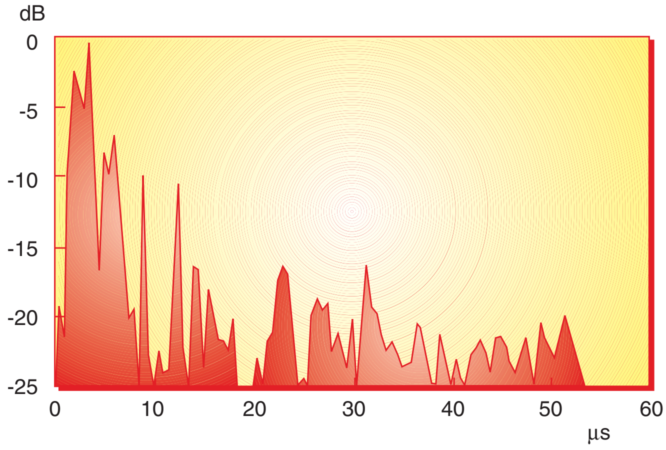


Figure 18 Typical IR from Farm_2, 953 MHz. Total received power was –93 dBm

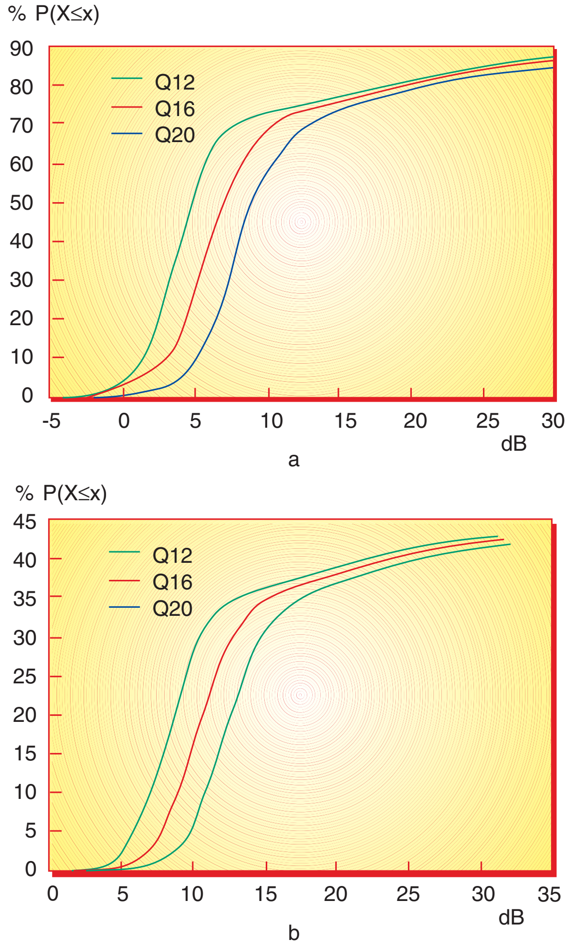


Figure 19 Cumulative distributions of averaged Q-parameters for Val_1
a) 953 MHz
b) 1718 MHz

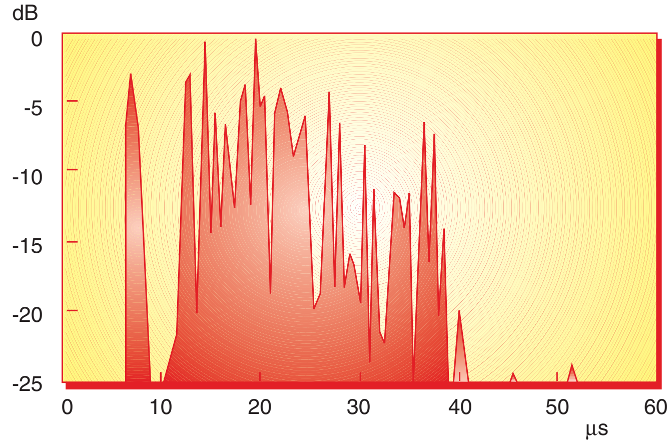


Figure 20 Typical IR from Val_1, 1718 MHz. Total received power was –72 dBm

Val_2: A rural area in a quite steep (angle 50 – 70˚) valley, 0.3 – 1 km wide surrounded by 700 m high mountain chains. Some forest in the more gentle hillsides. The transmitter antenna was placed in the bottom of the valley at road level on an 8 m high antenna mast. The measurement route followed the valley at a distance of 0.7 – 10 km from the TX. Half-way down the route the valley made a 60˚ turn, giving NLOS conditions further down the route. The upper part of the route had partly NLOS conditions due to obstacles close to the road. In general, we found the longest IRs in the NLOS area closest to the transmitter, and as we drove away from the transmitter, the IRs became shorter.

Q_{16} tells us that in this environment multipath propagation would cause no problem for GSM or DCS 1800.

4.3.3 Measurements in suburban terrain

Measurements have been conducted from one transmitter site in a suburban area of Oslo. The area is a smooth “U”-shaped valley, 3 km wide with approximately 200 m high surroundings. The hillsides were partly wooded, partly residential areas. The buildings were from 2 to 12 storeys high, mostly made of concrete. The TX was placed in the hillside 50 m above street level and was not an NMT base station site. The results were taken from two measurement routes. Due to GSM transmitters operating in the area, we were unable to perform 953 MHz measurements.

Sub_1: The distance to the transmitter was 0.1 – 2.5 km. Most of the route had LOS conditions, but temporary shadowing occurred when buildings were obstructing the direct signal path. The length of the route was 5 km.

Sub_2: The measurement route was 12 km long and formed a rectangular shape in the bottom of the valley. Distance from TX was 0.1 – 5 km. Mostly NLOS conditions due to shadowing from houses and small hills.

As we see from Figures 22 and 23, DCS 1800 will work satisfactorily in Sub_1, but not so well in Sub_2. This despite the fact that both routes are in the same environment. The main reason is of course that the former mainly had LOS whereas the latter mainly was an NLOS route. Hence, these routes clearly demonstrate the importance of careful base station site planning.

4.3.4 Measurements in urban areas

Urb_1: Urban environment with typically 4 – 7 storey buildings, mainly made of concrete. The antenna was placed below most roof tops in a small park, 13 m above street level. The measurement route was 4.2 km long with a distance 0.1 – 1.3 km from the TX, mainly with NLOS conditions. The street width varied between 4 and 10 m. The street pattern was not regular. We were unable to perform measurements at 953 MHz, due to several GSM base stations transmitting in the band used by our 900 MHz equipment. The received energy was never outside the sliding window length of 12 µs or larger. DCS 1800 would not suffer from intersymbol interference in this environment, and no Q-parameter curves are therefore presented.

4.4 Street descriptions for microcell measurements

We have concentrated on measurements in city streets and city squares, and have partly performed measurements in the same environments at 1950 MHz and 59 GHz. The description of the measurement scenarios for microcellular measurements at both frequencies will thus be given in this section.

In Figure 26 the measured city streets and city squares are shown. The receiver was moving, while the transmitter was stationary. The shaded areas in the figure represent buildings. In streets A to C there were 4 to 7 storey buildings, made of concrete. These buildings were built late in last century or in the beginning of this century. In street F there were modern buildings, made of concrete and steel, often covered with glass. The buildings on one side of street A were isolated buildings, separated by 5 – 10 metres, in all other streets there were no open spaces between the buildings. In all the streets cars were parked along one or both sides.

The two city squares are also shown in Figure 26. All shaded areas are concrete buildings, the building at the bottom of square I was covered with metal plates. Square J is two-levelled; the street in the bottom of the figure (level II) was about 8 metres above the main part of the square. Several cars were parked in both

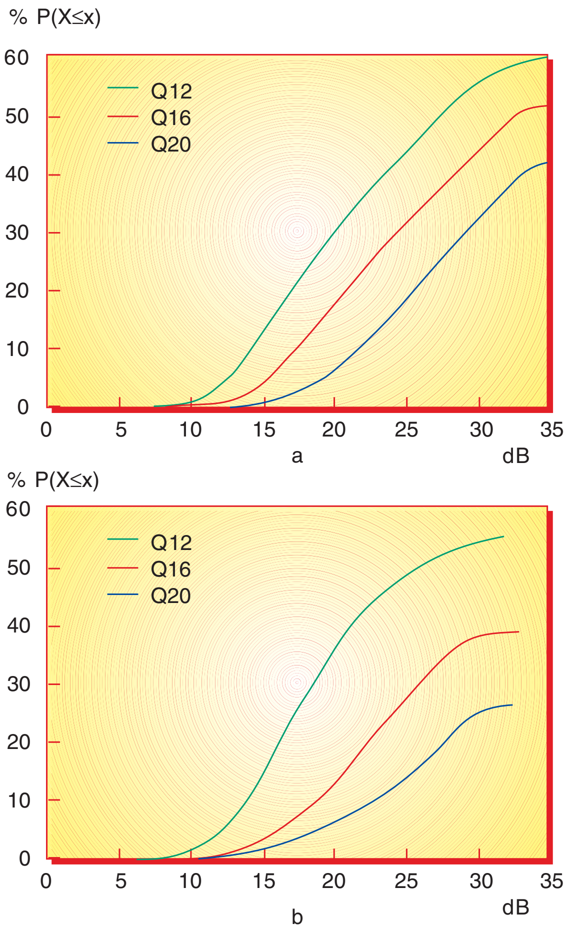


Figure 21 Cumulative distributions of averaged Q-parameters for Val_2
a) 953 MHz
b) 1718 MHz

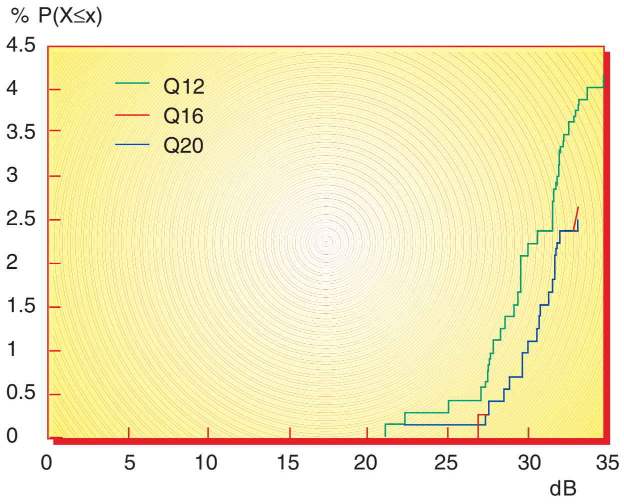


Figure 22 Cumulative distribution of averaged Q parameters for Sub_1, 1718 MHz

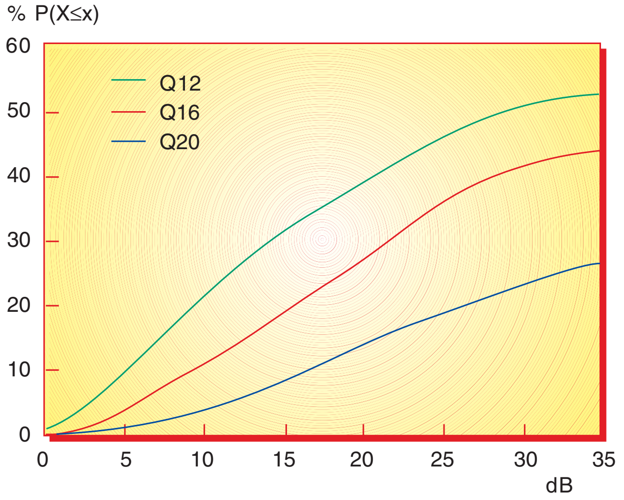


Figure 23 Cumulative distribution of averaged Q parameters for Sub_2, 1718 MHz

squares, only a few cars were moving. Transmitter antenna heights were 5 metres in the city square measurements.

The distance between each IR measurement was constant in each route, typically 0.5 metre. The height of the transmitter is indicated in Figure 26, the receiver height was 2.2 metres in all measurements. In all measurements the receiver was moving while the transmitter was stationary.

4.5 Multipath measurements in microcells at 1950 MHz

Due to the enormous growth in the number of subscribers to the mobile services, capacity problems are already experienced in the GSM system. One proposed solution to avoid capacity problems is to offer interworking between traditional mobile communication systems, like GSM, and cordless systems like DECT.

The idea is to build a microcellular DECT system to carry the traffic in areas where high capacity is needed and propagation conditions are such that they can be handled by a system employing a simple radio interface. On the other hand, where capacity demands are not so high, it is more feasible to let a macrocell based GSM system handle the mobile traffic. Work has already started to specify the interworking between DECT and GSM in a system employing dual-mode terminals.

From the description in Chapter 3.2.3 we remember that the DECT bit interval is 868 ns, and that a DECT implementation will fulfil the General Access Profile without employing a channel equaliser. It is, however, worth mentioning that some manufacturers offer DECT implementations employing antenna diversity to improve radio system performance.

We remember that for a radio system not employing a channel equaliser, the delay spread parameter gives an indication of the system performance. According to the rule of thumb for performance of such systems, DECT performance limit of bit error rate 10^{-3} will be reached at a delay spread of about one tenth of the DECT bit interval. The maximum tolerable delay spread for a DECT imple-

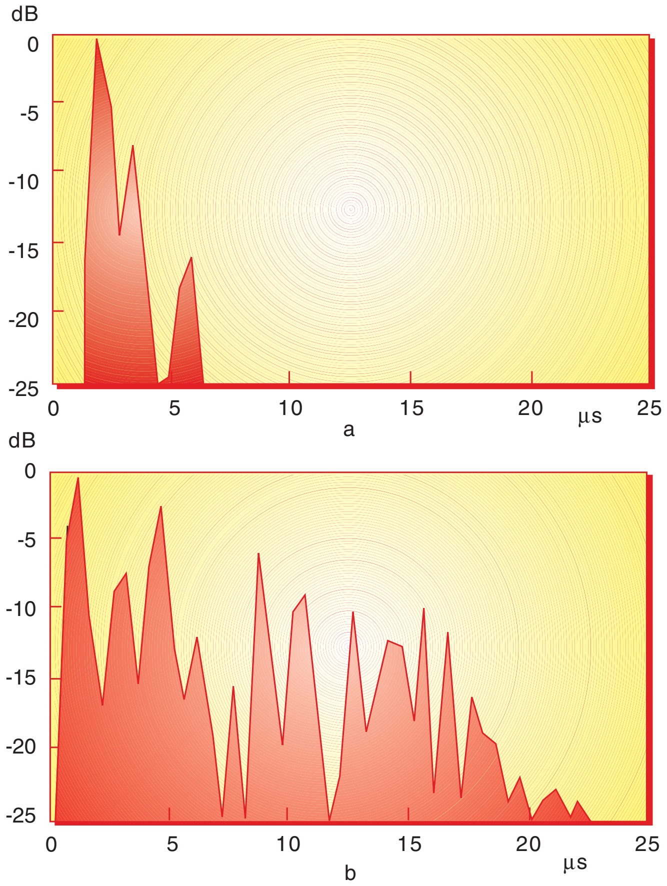


Figure 24 Two typical impulse responses from the suburban routes 1718 MHz. Total received power was –72 dBm and –92 dBm, respectively

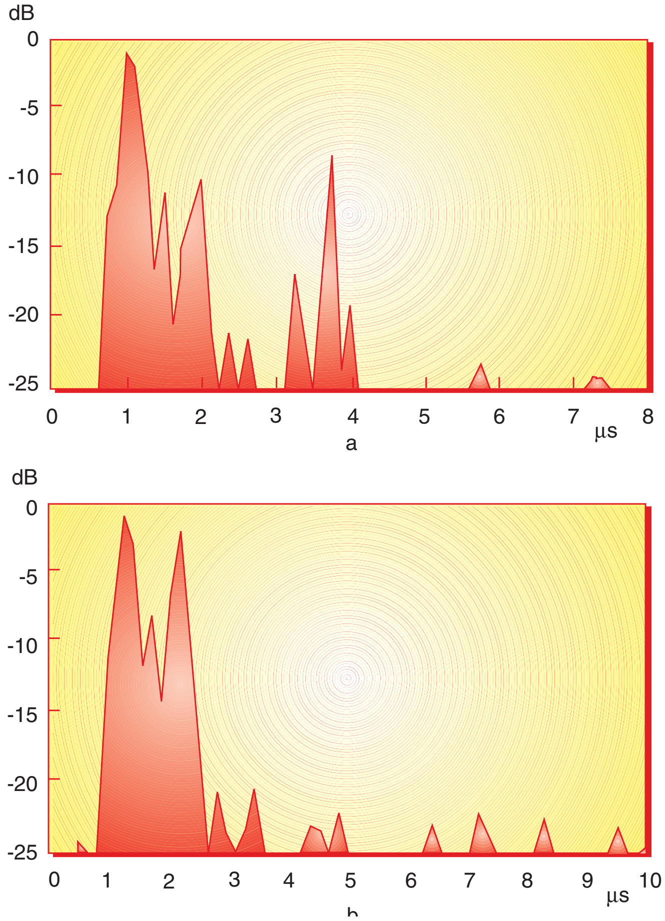


Figure 25 Two typical IRs from Urb_1, 1718 MHz. Total received power was –92 dBm in both cases

mentation not employing channel equalisation or a sophisticated antenna diversity algorithm is then below 100 nanoseconds.

If the delay spread is larger than this value, DECT reception without an equaliser would be limited by multipath propagation. The C/I requirement for DECT is 10 dB (91 % of energy within equaliser window). Hence, a DECT receiver equipped with a two-state Viterbi equaliser will work properly if the 90 % SDW is smaller than the DECT bit interval. Improvements are due to the fact that the Viterbi equaliser, employing maximum likelihood sequence estimation, is able to cope with intersymbol interference and at the same time utilising the multipath diversity (giving smaller fluctuations in the input power level).

4.5.1 Measurement results

All measurements reported here were performed with 20 ns resolution (50 MHz measurement BW). Output transmit power at 1950 MHz was 25 dBm, well aligned with the 250 mW output power of DECT. Omnidirectional λ/4 dipoles were used as transmitter and receiver antennas.

In this section measurements from two city streets downtown Oslo and from two city squares will be reported.

During measurements, the line-of-sight (LOS) path was never obstructed by large vehicles, but in street A trees sometimes blocked the LOS path.

Since the measurements mainly are aimed at a DECT implementation not employing channel equalisation, we have shown the cumulative distribution of delay spread from route A2 in Figure 27.

Looking at Table 3, we see that in all street measurements the delay spread is less than 65 ns in more than 90 % of the situations along each route. Only for a very small fraction of the city street measurements the delay spread exceeds one tenth of the symbol interval. Using the rule of thumb for irreducible bit error rate for GMSK modulation, we can therefore conclude that in city streets, multipath propagation is generally not limiting the DECT performance when the line-of-sight path exists.

From the non line-of-sight case we have no experimental results connected to multipath measurements, but experiences from an ongoing DECT field trial in Norway indicate that shadowing rather than

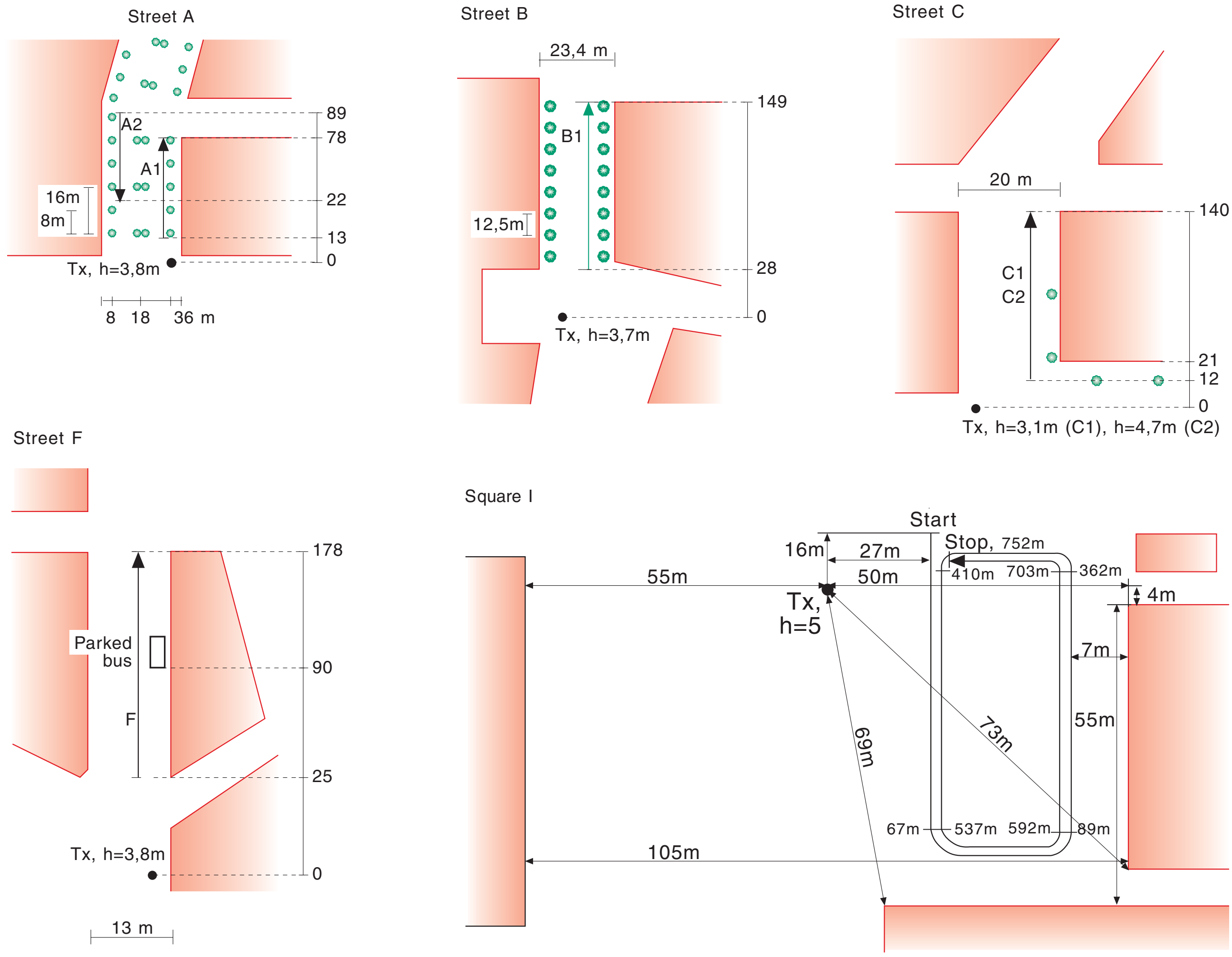
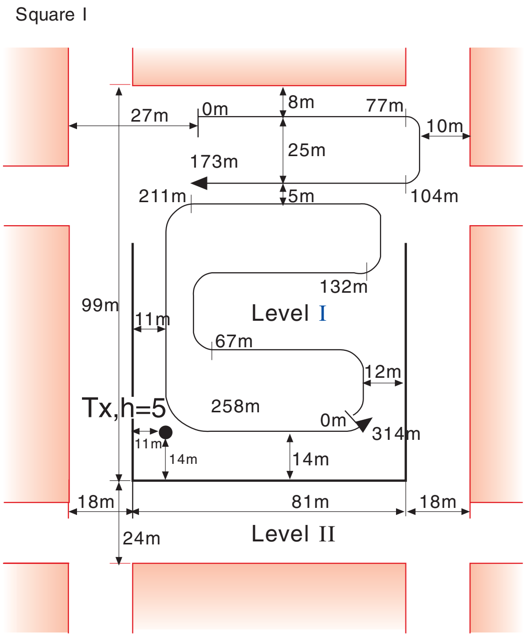


Figure 26 Transmitter positions (TX) and measurement routes in the measured streets and city squares. The shaded areas represent buildings, circles represent trees. All lengths along the streets and squares are in metres



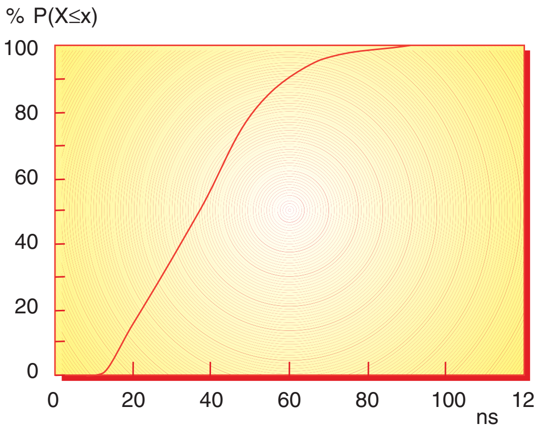


Figure 27 Cumulative distribution of delay spread, city street measurement A2

Table 3 Scenario description, delay spread and 90 % delay window from the two streets and two city squares. All parameter values in nanoseconds

Rte.	#	W [m]	Tr.	DS 50%	DS 90%	90% FDW 50%	90% FDW 90%	90% SDW 50%	90% SDW 90%
A1	1403	36	m	28.1	49.3	60	140	40	100
A2	1391	36	m	38.4	62.6	100	180	60	120
B1	2264	23	a	27.5	55.3	60	160	40	80
I	1829			53.4	102.6	120	300	60	160
JI	3956			93.8	145.1	260	420	140	340
J2	2000			98.8	148.2	280	420	180	320

Rte. = Route number
# = Number of IRs along the route
W = Width of street
Tr. = Trees: m: along and in the middle of street
a: along the street

DS = Delay spread
FDW = Fixed delay window
SDW = Sliding delay window

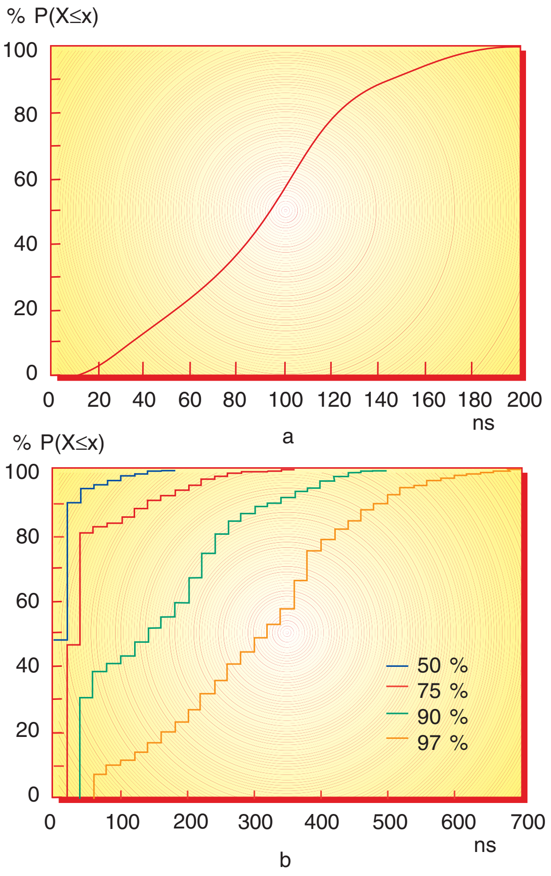


Figure 29 Cumulative distributions, city square measurement J1
a) delay spread
b) sliding delay window

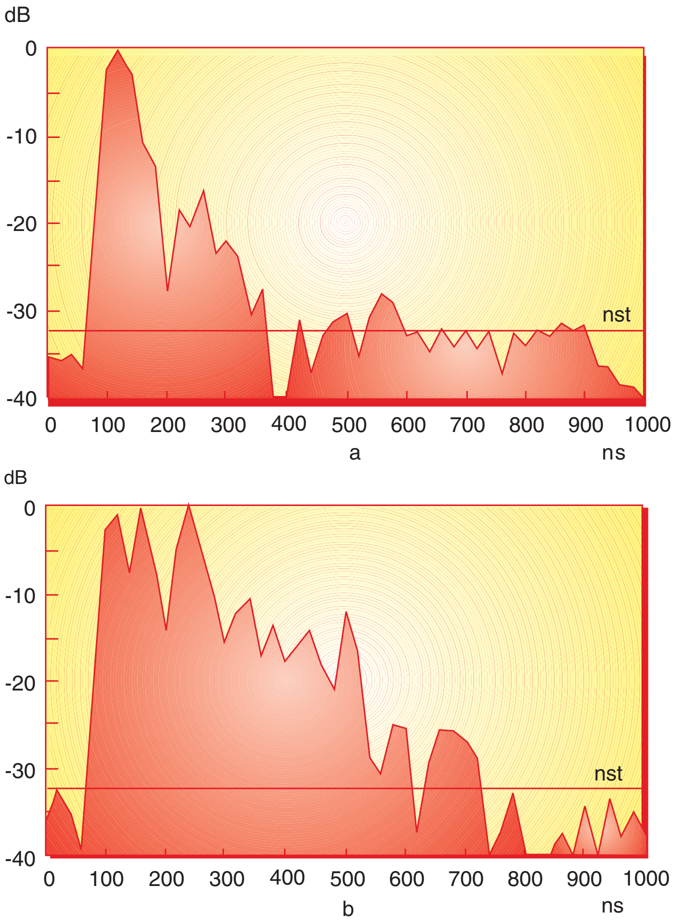


Figure 28 Two impulse responses from the city street measurements. Power levels are relative to peak

multipath propagation is limiting performance in those cases [13].

In Figure 28 two impulse responses are shown. They are measured in street A2, but give a good impression of typical situations in both streets.

Characteristic for the city squares compared to the streets is that the dimensions are larger. From the cumulative distribution of delay spread for measurement J1 and Table 3 we see that the larger dimensions also result in larger parameter values.

As is seen from Figure 29a and Table 3 the delay spread is less than 80 ns only in less than 40 % of the measurements from city square J. The rule of thumb for GMSK performance gives that in more than 60 % of the intended coverage area multipath propagation will cause severe problems to DECT communications. We also see from Figure 29b that the 90 % sliding delay window parameter (corresponding to C/I = 9.5 dB) is less than 340 ns in 90 % of the cases for all city square measurements. This clearly demonstrates that any means introduced in DECT to make the system capable of coping

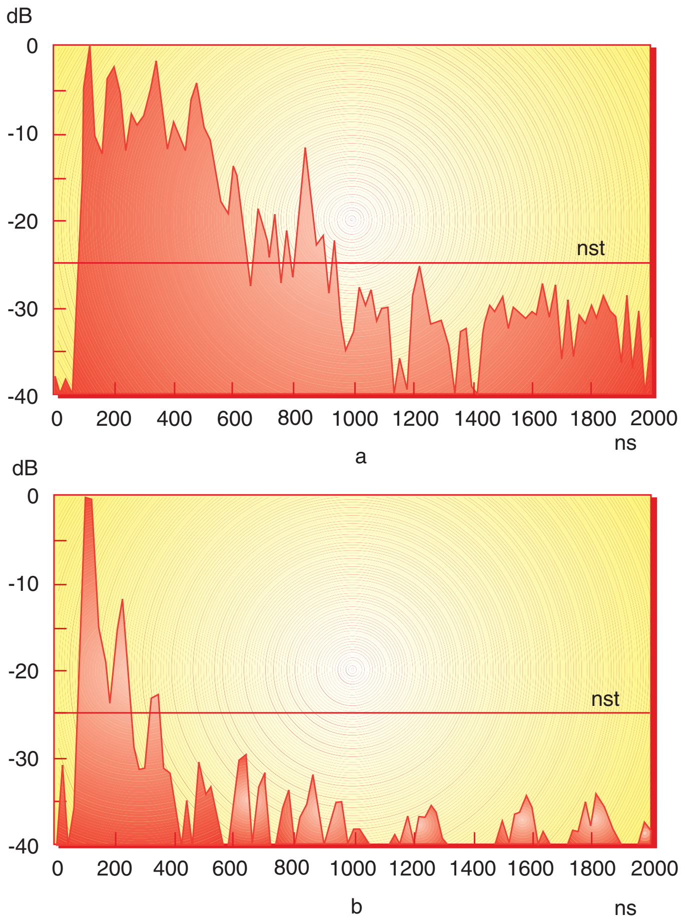


Figure 30 Two impulse responses from the city square measurements. Power levels are relative to peak

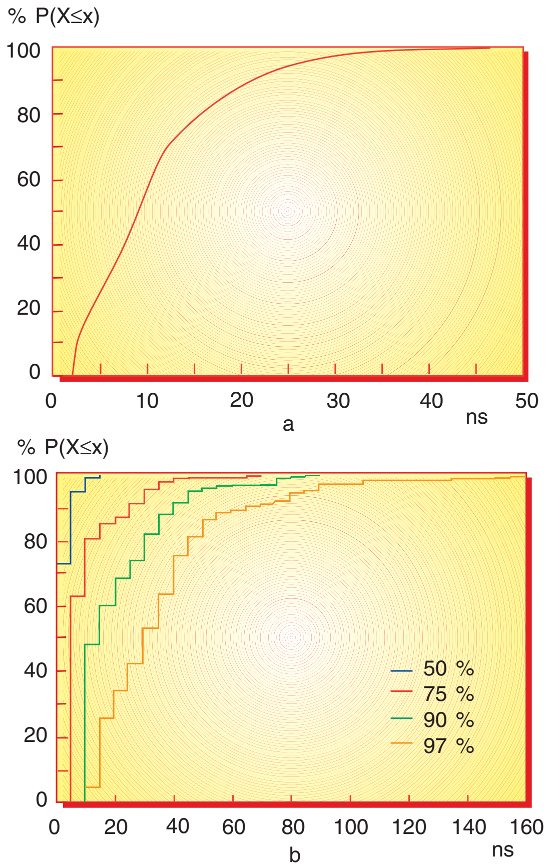


Figure 31 Cumulative distributions for city street measurement 2A
a) delay spread
b) sliding delay windows

with multipath stretching over some hundreds of nanoseconds significantly would improve system performance.

Two impulse responses from city square measurement J1 are shown in Figure 5. These impulse responses can be regarded as typical for all the city squares.

4.6 Multipath measurements in microcells at 59 GHz

The U-band[3] is intended for use in mobile broadband services, amongst others. In the European research project RACE Mobile Broadband System (MBS), studies of such a microcellular high-capacity system aiming at providing B-ISDN services to moving and movable users are performed [14]. Considerations are given to the frequency bands 62 – 63 GHz and 65 – 66 GHz, in addition to a band of 2 GHz width around 40 GHz. User bit rates may be as high as 155 Mb/s. Hence, the system bandwidth will be of the same order of size as the maximum measurement bandwidth of Telenor R&D's channel sounder.

Here we present measurements from four downtown Oslo streets and from two city squares. Transmitter sites and measurement routes from city street measurements A and B and city square measurements I and J were identical to those corresponding to 1.95 GHz measurements labelled A, B, I and J. A more complete description of the measurement scenarios and a larger set of results are given in [15].

Only instantaneous IRs are treated, i.e. the receiver moved only a fraction of a wavelength during each IR measurement, and no averaging to remove fast fading on each tap of the IR was performed.

4.6.1 Measurement results

All measurements reported in this chapter were performed with 5 ns resolution (200 MHz BW). Output transmit power was 500 mW. Directive transmitting antenna (90 degrees horizontal beam width) and omnidirectional receiving antenna were used during measurements at 59 GHz. During city street measurements, the transmitting antenna was pointing down the measured street. For the city square measurements, antenna pointing directions are shown in Figure 26.

During measurements, the line-of-sight (LOS) path was never obstructed by large vehicles, but in street A trees sometimes blocked the LOS path. The weather during all measurements was sunny, and street surfaces were dry.

Cumulative distributions of the delay spread and sliding delay windows for a

[3] *The frequency band 40 – 60 GHz is called the U-band.*

Table 4 Scenario description, delay spread and 90 % delay windows from the four streets and the two city squares. All parameter values in nanoseconds

Rte.	#	W [m]	Tr.	Cr.	DS 50%	DS 90%	90% FDW 50%	90% FDW 90%	90% SDW 50%	90% SDW 90%
A1.w	148	36	m	s	5.3	14.0	5	25	10	10
A1.s	148	36	m	s	7.6	16.5	20	35	10	25
A2.w	144	36	m	s	12.5	27.9	35	70	25	45
A2.s	124	36	m	s	12.7	22.7	40	70	30	45
B1.w	303	23	a	s	4.1	8.6	10	20	10	15
B1.s	273	23	a	s	4.2	6.5	10	20	10	20
B2.w	211	23	a	s	3.4	8.7	10	25	10	15
B2.s	153	23	a	s	3.3	6.0	10	20	10	15
C1.w	326	20	n	n	8.4	13.1	20	40	15	30
C2.w	328	20	n	n	8.5	13.7	20	40	15	30
F.w	315	13	n	m	3.6	48.3	10	80	10	40
I.s	1238				28.1	76.6	55	195	15	150
J1.s	610				40.8	92.2	45	270	10	135
J2.s	483				37.8	83.7	90	220	15	160

Rte. = Route name
# = Number of IRs along the route
W = Width of street
Tr. = Trees: m: along and in the middle of street
a: along the street
n: no trees
Cr. = Moving cars: f: only a few
s: some
m: many
n: none

DS = Delay Spread
FDW = Fixed Delay Window
SDW = Sliding Delay Window

measurement scenario in street A are shown in Figure 31.

In Table 4 the 50 % and 90 % values of delay spread and fixed and sliding delay windows are shown for the city street and city square measurements. The measurements marked “w” in Table 4 were conducted in March, hence the trees had not come into leaf. The measurements marked “s” were conducted in June, with trees in leaf.

In Figure 32 two typical impulse responses are shown. They are measured in street A, but give a good impression of typical situations in all streets.

Many of the measured IRs looked like the upper in Figure 32, i.e. with only one peak. This peak, however, may consist of several rays. The IR resolution of 5 ns, corresponding to a spatial resolution of 1.5 metre, is not able to resolve all first order reflections from the buildings and street surface, especially when the distance between transmitter and receiver is large.

In all streets except street F the delay spread is less than 20 ns in more than 90 % of the situations along each route. The 90 % fixed delay window is 50 ns or less in 90 % or more of all measurements in each route, except the route in street F.

The reason for the higher values in street F is a tourist coach parked as shown in Figure 26. This coach created a stable reflection, which can be seen on the impulse response shown in Figure 33a. In Figure 33b the 97 % fixed delay window as function of measurement number is shown, and we see clearly that the main contribution to the high FDW values is a stable stationary reflection source which is passed by the receiver around measurement number 160. This corresponds to the position of the coach.

In general, if the streets are empty (no major reflection sources) all parameter values will increase with increasing street widths. In street B, which is a street with trees along the buildings on each side, we observe parameter values of equal or less magnitude compared to street E, which is a narrower street. This is probably because the trees reduce the reflections from the walls.

In all measurements we started measuring when entering or just before entering the main lobe of the transmitter antenna (i.e. the transmitter antenna was pointing towards the receiver). We did however notice that when the distance between transmitter and receiver was very short (5 – 10 metres) and the transmitter did not point to the receiver, bad multipath situations often occurred. This is because in such situations the possible excess delays are maximised, and since the direct signal is attenuated due to the transmitter antenna pattern, the LOS ray may be of the same order of magnitude as the reflected ones. It is important to be aware of this dangerous situation, because often the antenna patterns are chosen to get the received power as uniform as possible in the cell, i.e. radiating more power to the distant parts of the cell than to the part below the base station.

Finally, looking at the received power, the observed fades measured with 200 MHz bandwidth were typically 3 – 6 dB. (Narrowband measurements showed typical fades of 15 – 20 dB, sometimes deeper.) When the LOS path was obstructed by a tree, as in route A2, this resulted in a fade of depth 8 – 12 dB, when measuring both narrowband and wideband.

4.7 Comparison of propagation conditions at 1950 MHz and 59 GHz

1950 MHz and 59 GHz measurement campaigns performed in identical environments make it possible to compare statistics describing multipath conditions in the two frequency bands. We see from Tables 3 and 4 that values of parameters describing the multipath situation are significantly larger at 1950 MHz than at 59 GHz. This is mainly due to two reasons:

- influence of transmitting antenna patterns
- different reflection properties at the two frequencies.

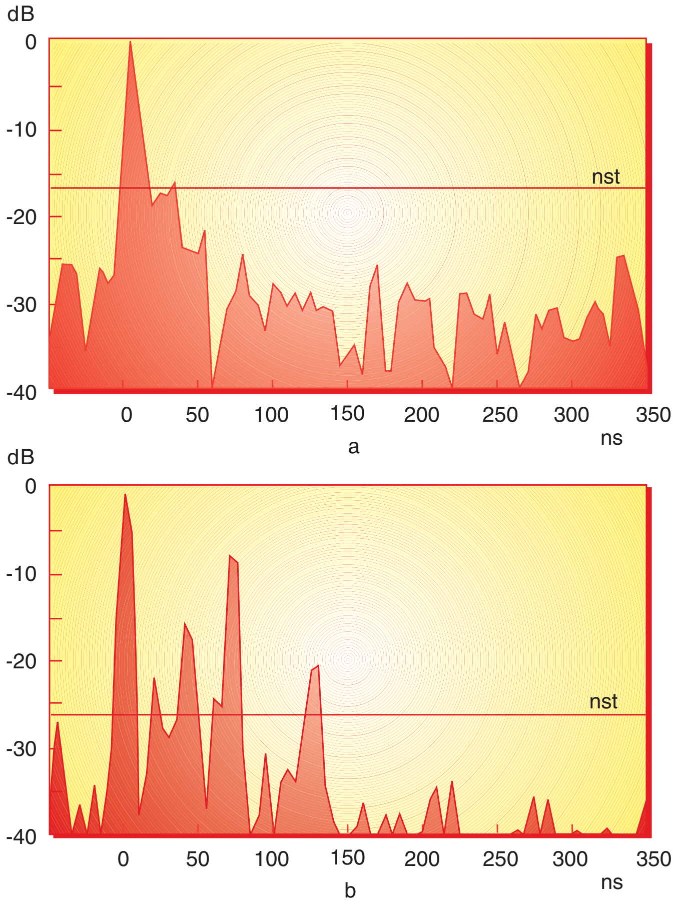


Figure 32 Typical impulse responses for all city street measurements. IRs are recorded along measurement route A2. Power levels are relative to peak

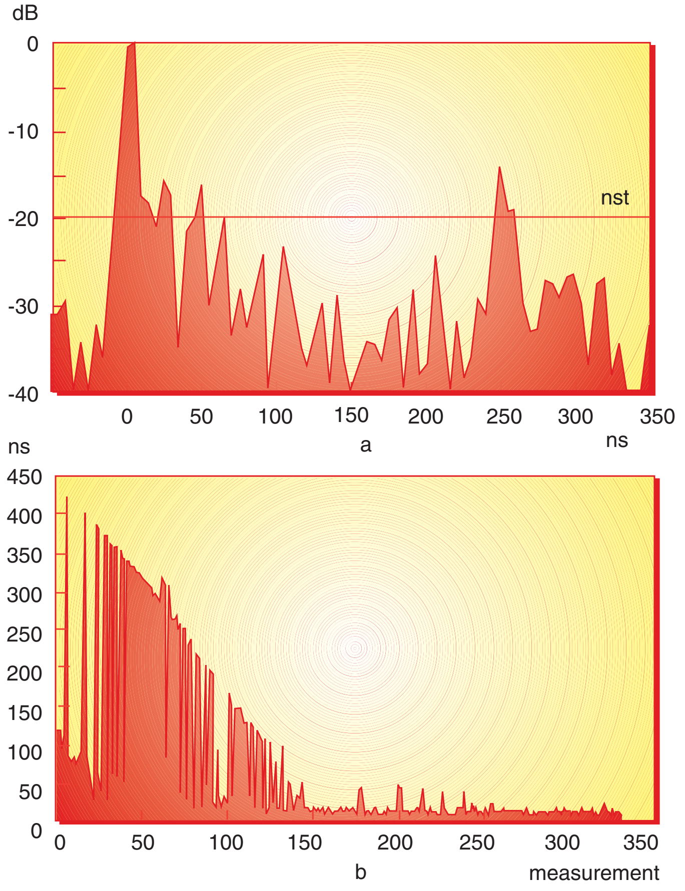


Figure 33 Impulse response and 97 % fixed delay window as function of measurement number from city street F. The coach parked in the street creates a stable reflection source

During all measurements at 1950 MHz, omnidirectional transmitting antenna was used, whereas directional transmitting antenna with half power beam width of 90 degrees in the horizontal plane was used at 59 GHz. Use of directional transmitting antenna reduces the magnitude of reflected signal components originating from reflecting objects situated within the back lobe of the transmitting antenna. Reduced magnitude of those reflected signal components compared to the direct one is expected to reduce problems due to multipath propagation.

Another effect is originating from the different carrier frequencies. At 1950 MHz, the wavelength is 15.4 cm. At 59 GHz, the wavelength is just above 5 mm, hence being comparable with the roughness of many surfaces. Wavelength comparable with surface roughness gives rise to diffuse rather than specular reflections, giving smaller reflected components.

In addition, there is an extra signal attenuation at 59 GHz due to the oxygen absorption in this frequency band. Signal components with long excess path lengths are very much attenuated due to the extra signal attenuation of about 15 dB/km. This effect is also reducing multipath problems at 59 GHz.

All these factors would contribute to smaller parameter values at 59 GHz than at lower frequencies. This is well in line with results shown in Tables 3 and 4.

All these factors tend to give larger multipath spread at UHF than in the U-band, and we have not been able to isolate the effects of the differences in transmitting antenna patterns, oxygen absorption and wavelengths.

5 Conclusions

The theory behind the mobile radio propagation channel has been presented. The radio channel as a function of time and position has been explained in some detail.

Then different principles of wideband radio channel measurements, channel soundings, have been introduced. An overview of Telenor R&D's channel sounders for use in the UHF- and U-bands have been given, together with the parameters used to describe the multipath situation.

Parallel multipath and GSM measurements conducted by Telenor R&D show that the Q_{16}-parameter, derived from multipath measurements, gives a correct estimate of the GSM signal quality in more than 90 % of the cases.

Parallel multipath and DECT measurements show that the delay spread parameter is an appropriate link quality estimator for a DECT implementation not employing channel equalisation.

Lastly, measurements performed at 900 and 1700 MHz in different macrocellular environments in Norway have been presented, together with measurements from microcellular environments at 1950 MHz and 59 GHz.

We have demonstrated that due to multipath propagation, there is no guarantee that a base station site suitable for an analogue radio communication system is a good choice for a digital system. Even radio systems with very advanced air interfaces, like GSM and DCS 1800, may experience problems due to multipath propagation if the base station sites are not chosen properly.

We have also performed measurements in microcellular environments, aimed at DECT and the future Mobile Broadband System.

In city streets DECT link quality is normally not limited by multipath propagation. In city squares, however, where the dimensions are larger, multipath propagation would often cause severe problems to DECT communications, raising the need for advanced diversity techniques or channel equalising.

At 59 GHz, the wavelength is comparable with the roughness of many surfaces. This gives rise to diffuse rather than specular reflections, and effects of multipath propagation are thus smaller than at lower frequencies. Bearing in mind the proposed bit rates for the MBS and related systems, multipath propagation may still be a limiting factor in many environments for the real broadband radio communications services.

6 References

1 Calhoun, G. *Digital cellular radio.* Norwood, MA, Artech House, 1988. ISBN 0-89006-266-8.

2 Lee, W C Y. *Mobile communications engineering.* New York, McGraw-Hill, 1982. ISBN 0-07-037039-7.

3 Jakes, W C. *Microwave mobile communications.* New York, Wiley, 1974. ISBN 0-471-43720-4.

4 COST. *A note on definitions of terms for impulse responses.* Oct. 1989. (COST 231 TD(89)060.)

5 Løvnes, G, Paulsen, S E, Rækken, R H. *UHF radio channel characteristics. Part one: multipath propagation and description of the NTR channel sounder.* Kjeller, Norwegian Telecom Research, 1991. (TF-report R 54/91.) ISBN 82-423-0172-7.

6 Løvnes, G, Paulsen, S E, Rækken, R H. *UHF radio channel characteristics. Part two: wideband propagation measurements in large cells.* Kjeller, Norwegian Telecom Research, 1992. (TF-report R 19/92.) ISBN 82-423-0214-6.

7 Trandem, O et al. *Equipment for multipath measurements at 60 GHz : specifications version 3.* Trondheim, SINTEF DELAB, 1992. (SINTEF DELAB document 40-NO920378.) (Restricted; in Norwegian.)

8 Trandem, O et al. *Channel measurements at 60 GHz : first measurement of time consumption in the DSP program.* Trondheim, SINTEF DELAB, 1993. (SINTEF DELAB document 40-NO930086.) (Restricted; in Norwegian.)

9 Rækken, R H. *Description of the DSP and PC software used with NTR's 60 GHz channel sounder.* 1993. (RACE document MBS/WP4.3.1/ NTR019.12.) Also: Kjeller, Norwegian Telecom Research, 1993. (TF-report N 35/93.)

10 Løvnes G, Paulsen, S E, Rækken, R H. Multipath measurements for GSM and DCS 1800. In: *Fifth Nordic seminar on digital mobile communications (DMR V),* Helsinki, 1992, 347–354.

11 Løvnes, G, Paulsen, S E, Rækken, R H. Estimating GSM coverage using 900 MHz multipath measurements. In: *44th Vehicular Technology Conference,* Stockholm, June 1994, 1798–1802.

12 Rækken, R H, Eskedal, B E. DECT performance in multipath environments. In: *Nordic radio symposium 95,* Saltsjöbaden, April 1995, 133–138.

13 Eskedal, B E. DECT field trial at Førde : examining the performance of DECT/Ericsson in a multi-operating environment. In: *IIR conference on developing and exploiting wireless local loop,* London, May 1995.

14 Fernandes, L. Overview of the project R2067. In: *RACE Mobile Telecommunications Workshop.* Metz, June 1993, 75–79.

15 Antonsen, E et al. *59 GHz wideband propagation measurements.* Kjeller, Norwegian Telecom Research, 1994. (TF-report R 36/94.) ISBN 82-423-0304-5.

UMTS – The Universal Mobile Telecommunications System

BY DAG FREDRIK BJØRNLAND AND GEIR OLAV LAURITZEN

1 Introduction

The telecommunication market today is to a large extent dominated by two trends:

- Increased demand for communications on the move
- Increased demands for higher bitrate services (e.g. multimedia services and data services).

The coupling of these two trends will create a mass market demand necessitating the development of a new generation of mobile telecommunication systems emphasising spectrum-efficiency, enhanced mobility support and the support of broadband services and applications.

1.1 History

Within the ITU[1] initial studies on the new generation mobile system concept, *Future Public Land Mobile Telecommunications Systems (FPLMTS),* were initiated in 1986. After several years of feasibility studies, it was recommended during the 1992 World Administrative Radio Conference (WARC'92) [1] that 230 MHz of spectrum in the 2 GHz band should be reserved for FPLMTS. The 230 MHz band designation by WARC'92 boosted the further progress of standards work relating to third generation mobile telecommunications systems.

The ETSI[2] standards work on a European implementation of FPLMTS, *the Universal Mobile Telecommunications System (UMTS),* was initiated by the Technical Committee Special Mobile Group (TC SMG) in late 1991. SMG set up sub-technical committee SMG5 to carry out the task as system architect for UMTS. Since then, SMG5 has distributed most of the technical work to other sub-technical committees in ETSI, including SMG1 WPC (UMTS service aspects), SMG3 WPD (UMTS network aspects) and SMG SEG (UMTS security aspects). The radio aspects are still handled by SMG5 but are likely to be moved to SMG2 during 1996.

SMG5 currently meets four weeks per year in addition to a large number of ad hoc meetings, enjoying the regular participation of 60 – 80 professionals from manufacturers and operators from most European countries. SMG5 is currently chaired by Mr. Juha Rapeli, Nokia Mobile Phones, Finland.

1.2 The UMTS standardisation process

The UMTS standardisation process can be split into three main phases [2]:

1 *The research phase,* where the main focus is on technological research

2 *The concept phase,* where focus is on objectives, requirements and framework for the system.

3 *The specification phase,* which establishes the specifications necessary to meet the objectives and requirements.

The first two phases are expected to be concluded for most areas by the end of 1995, and UMTS is now in the process of moving into the specification phase. Expected finalisation of the specifications will be towards the end of 1998, in time to meet the current target date for the introduction of UMTS: *1 January, 2002.*

2 UMTS key objectives

The general objective of UMTS is to provide global high quality services to mobile users using the WARC'92 identified frequencies. An additional objective is to support *personal communication services (PCS),* that is, to enable the user to access whatever services he wants, when he wants and in whatever environment he finds himself. The concept of PCS for UMTS is illustrated in Figure 1.

More detailed objectives can be found in [3], and include:

- *Integration of Services and Application Areas:* UMTS will offer support in one system for application areas and services currently provided by dedicated systems. Target application areas and services include paging, mobility services (terminal and personal mobility, service mobility), mobile data, mobile telephony, public cellular applications, private business applications, residential cordless applications, wireless PABX applications and mobile satellite access.
- *Integration of fixed and mobile networks,* through the integration of fixed and mobile system technologies.
- *High Quality of Service,* comparable to that of the fixed network.
- *Services requiring a range of bit rates* (up to 2 Mb/s in phase 1) and variable bit rate services, supporting also data services and the emerging multimedia services.
- *Global terminal roaming capability,* enabling a user to access UMTS services in all regions of the world via other networks than his home network.
- *Satellite and terrestrial based coverage,* supporting direct access to the satellite com-

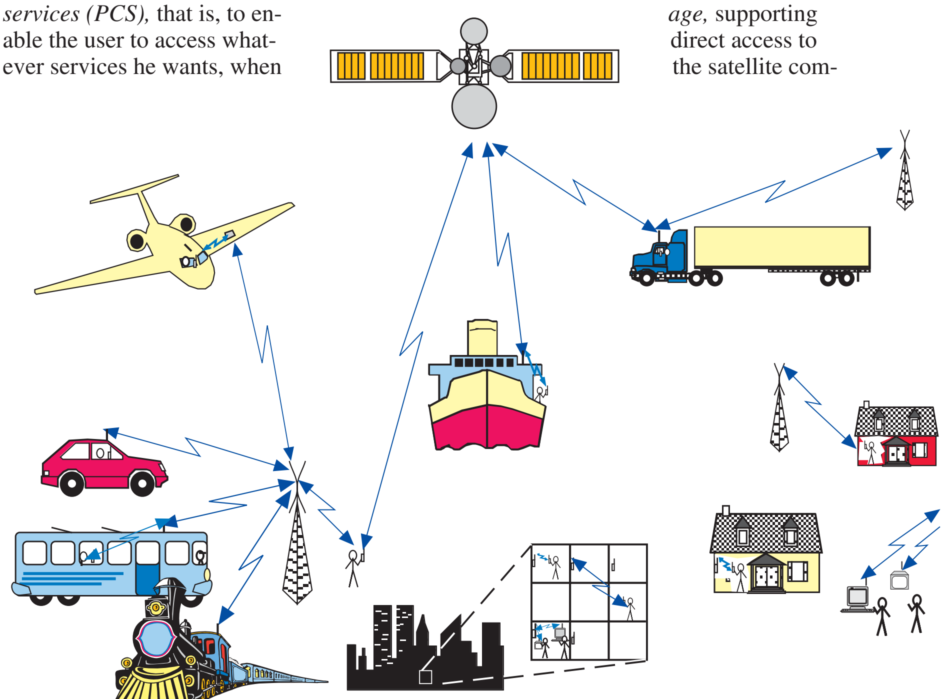


Figure 1 UMTS: Communications anytime, anywhere

[1] *The International Telecommunications Union*

[2] *The European Telecommunications Standards Institute*

ponent as a complement to terrestrial access. The satellite component will extend UMTS coverage to areas where it is not techno-economically feasible to provide terrestrial coverage (e.g. remote areas, aeronautical and maritime areas) and support global roaming.

3 UMTS system overview

3.1 UMTS service aspects

The UMTS role model

The UMTS role model defines the four main parties involved in the provisioning of UMTS services [4].

The four parties, or roles, defined in the model are:

- *The network operator,* which is the party responsible for the deployment and management of the transmission network
- *The service provider,* which is the party responsible for offering an agreed set of services to the customer
- *The subscriber,* which is a legal entity or an individual who subscribes to UMTS services through a UMTS service provider
- *The user,* which is a person, entity or process using UMTS services.

Key aspects of the role model are the separation of the UMTS user and the UMTS subscriber, enabling multiple users to be supported through a single subscription and billing arrangement, and the separation of service provider and network operator. The separation of service provider and network operator constitutes a key feature of the liberalised telecommunications market envisaged after 1998, and enables third party service providers equal access to UMTS networks. The network operator role can be further subdivided into a core and an access network operator, enabling independent access operators connection to a UMTS backbone (core) network in support of the concept of *Open Network Provision (ONP).*

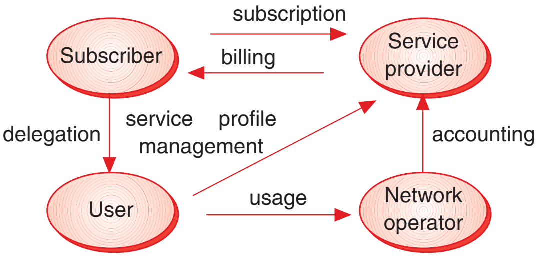


Figure 2 The UMTS Role Model

Table 1 UMTS bearer service classes

Class	A	B	C	D
Timing relation (source to destination)	Required		Not required	
Bit rate	Constant	Variable		
Connection mode	Connection oriented			Connectionless

The integration of service environments

UMTS is able to offer services through a *single* user equipment in public, private and residential service environments. Multi-mode terminals capable of supporting multiple UMTS radio interfaces and dual-mode terminals supporting both the terrestrial and the satellite component will further facilitate the integrated support of multiple service environments by a single terminal.

UMTS user services

To facilitate competition between service providers, UMTS teleservices, supplementary services and applications will not be standardised. Instead, the standardisation effort has been concentrated on standardising *service capabilities*. In particular, the support for ISDN bearer services and B-ISDN bearer services at least up to 2 Mb/s transfer rate has been recognised as an important goal for UMTS [5]. UMTS services will therefore be based on the bearer service classes defined for B-ISDN, as shown in Table 1.

UMTS mobility services

UMTS will support both *Terminal Mobility* and *Personal Mobility*. Personal mobility includes *Universal Personal Telecommunication (UPT)* and *UMTS User Mobility (UUM)*. UUM is distinguished from UPT mainly by being restricted to the UMTS boundary. UUM facilitates multiple user registrations on a single terminal (e.g. an office fax) and the registration by a single user for different services on different terminals (e.g. phone calls on one terminal and faxes on another).

UPT support in UMTS is limited to UPT registrations to terminals where a UMTS user is already registered. As UPT is a service also offered in the fixed networks, the support of UPT will facilitate the migration of users between mobile and fixed networks.

In addition to terminal and personal mobility, UMTS will also provide standard terminal portability (allowing standard terminal equipment like printers or fax machines to be connected directly to the mobile terminal via one or more standardised interfaces) and service mobility (allowing a roaming user access to his own set of personalised services in all networks covered by his subscription).

The relationships between the different types of mobility are illustrated in Figure 3.

3.2 UMTS network aspects

IN-based service provisioning

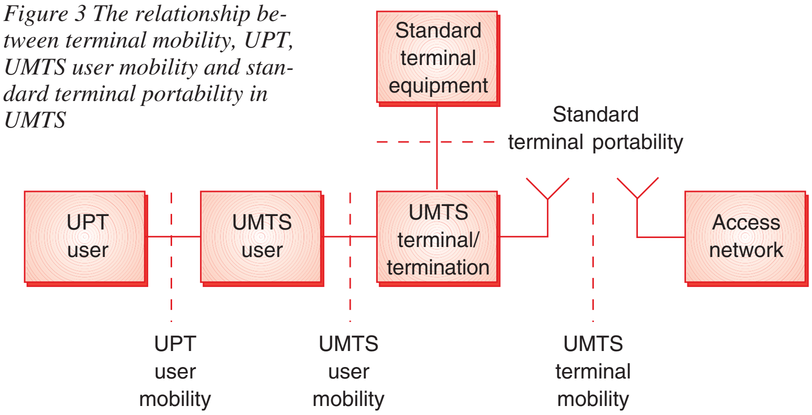
To support rapid and cost-effective service creation and the introduction of personalised services, UMTS is being developed using the *Intelligent Network (IN)* methodology, as defined in the ITU Q.1200-series. The IN concept supports the separation of the service provider and network operator *roles* through the separation of service creation and service control from the basic underlying call control. The support of third-party service providers through standardised interfaces to the underlying network, however, is still a subject for discussions in the fora standardising IN.

In light of the increasingly important role IN plays in the existing telecommunications networks and IN's expected role as the main platform for service provision in the fixed networks of the near future, choosing IN for UMTS service provisioning can be seen to facilitate the introduction of UMTS in the fixed networks.

Mobility management

The IN methodology has also been chosen as the basis for UMTS mobility management. This reflects the view that mobility can be considered as a set of services which are supplementing the functionality of the basic telecommunica-

Figure 3 The relationship between terminal mobility, UPT, UMTS user mobility and standard terminal portability in UMTS



tion services in networks for fixed communication. In fact, the analogy with traditional services can be taken even further, by observing that within the UMTS management framework [6], provisions have been made for the possible charging of the usage of the UMTS mobility services (specifically location management and handover). For example, a user frequently crossing cell boundaries could be charged explicitly for the extra signalling this produces.

UMTS network design methodology

The methodology used so far for UMTS network design has to a large extent been derived within the RACE project 2066 MONET [7]. This methodology enables the design in a step by step fashion, each step taking different design concerns into account. The methodology is based on the CCITT 3-stage methodology [8], and consists of four steps:

1 Operational and functional requirements are derived from user needs. These requirements lead to a precise description of the service the UMTS system must offer to its users.

2 In a subsequent step, service specific *Functional Models (FM)* are constructed. FMs consist of *Functional Entities (FEs)* and the relations between them. An FE consists of a subset of the functions needed to offer a particular service. Functions grouped within a single FE will always be realised in a single piece of equipment. In order to support their joint operation, interactions between FEs are described by means of *Information Flows (IFs)*.

3a Using the service-specific FMs, a *Functional Architecture (FA)* is derived. The FA is abstracted from the lower layer protocols that provide the means of reliable communication between application processes and facilitates the development of Application Layer protocols. The FA identifies *Network Entities (NEs)* that are the smallest entities that can exist in possible system implementations, and the *Functional Interfaces (FIs)* between them. The information exchange between the concerned FEs on the FI defines the information transfer requirements for an application layer protocol.

3b In order to identify the Network Entities, *Reference Configurations (RCs)* play an assistant role. A reference configuration shows possible network arrangements that may occur in system implementations.

4 Finally, the *Network Architecture (NA)* is defined. The NA also identifies the smallest components that can exist in possible system implementations, but this model is focused on the lower layer protocols. The NA consists of *Physical Entities (PEs)* and *Physical Interfaces (PIs)* between them. A PI is defined by the entire protocol stack and supports one or more FIs.

With respect to the CCITT 3-stage method, step 1 corresponds to Stage 1, step 2 corresponds to Stages 2.1 – 2.4 and Step 3 and 4 correspond to Stages 2.5 and 3.

Current work on UMTS network aspects focuses on step 2.

Functional models

A number of service specific functional models (FMs) and one generic functional model (G-FM) incorporating all the service-specific FMs have been proposed for UMTS. The generic functional model is shown in Figure 4 [9].

The G-FM is based on the IN CS-1 FM [10], containing the SDF, SCF and SSF/CCF FEs, which has been enhanced to meet UMTS specific requirements. In particular:

- The role of the *Service Data Function (SDF)* has been enhanced to also handle storage, access and maintenance of mobility related data, and to provide information for the SCF on routing (exit point from the core network), access rights and user information relating to handover.
- The role of the *Service Control Function (SCF)* has been enhanced to also contain the overall mobility control logic required to support all mobile specific functions (e.g. paging control, location and identity management and routing), and security management (e.g. authentication and ciphering management). Core network supervision for the handover procedure when handover involves multiple RACFs, possibly belonging to different administrative domains has also been added to the SCF. Service logic in the SCF can be invoked by service requests from other FEs to support location management and mobility management.
- The *Service Switching Function (SSF)* has been enhanced with a state machine to monitor bearer states within the SSF/CCF towards the SCF in a similar way the Basic Call State Model (BCSM) models the state of calls to the SCF.
- The *Call Control Function (CCF)* has been enhanced with functionality required for the radio access. This includes the execution of some types of handover. In addition, the fixed

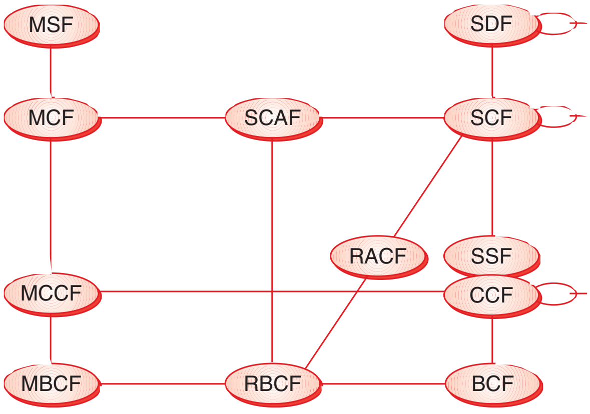


Figure 4 The UMTS generic functional model

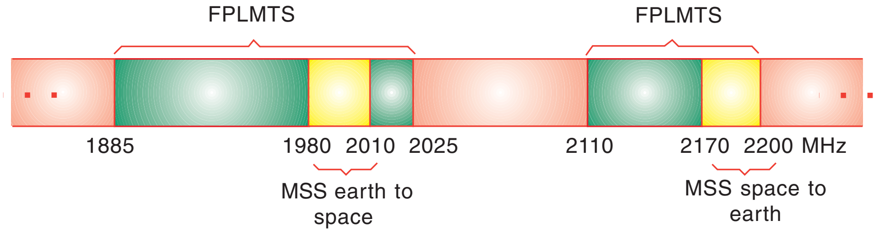


Figure 5 Spectrum identified for FPLMTS

bearer control functionality contained in the ISDN CCF has been separated out into a separate *Bearer Control Function (BCF)*.

In addition, some new FEs have been introduced on the fixed side of the radio interface. These are:

- *The Bearer Control Function (BCF)*, which has been introduced to control the bearer connection elements in the fixed network in order to provide the bearer capabilities requested by the CCF. The separation of bearer and call control (and of bearers and calls) is a fundamental requirement of the radio access of UMTS in order to provide efficient radio resource usage, to support macro diversity and to facilitate handover.
- *The Radio Bearer Control Function (RBCF)*, which has been introduced in order to control the radio bearer connection in the fixed part of the radio access sub-network.
- *The Radio Associated Control Function (RACF)*, which has been introduced to provide the control functionalities required in the radio access subsystem. The role of the RACF is to drive the actions which have to be performed by other FEs for radio resource allocation/deallocation and radio bearer connection establishment/release.
- *The Service Control Agent Function (SCAF)*, which has been introduced to provide a mechanism to invoke service logic without establishing a call (i.e. without using the CCF/SSF – SCF relationship). This applies in particular to the mobility management services. The SCAF is also involved in the paging procedure.

On the mobile side, a number of new FEs have also been added to the IN model as it was considered advantageous from a modelling perspective, to show peer-to-peer relationships across the air interface as well. The new FEs are:

- *The Mobile Storage Function (MSF)*
- *The Mobile Control Function (MCF)*
- *The Mobile Call Control Function (MCCF)*
- *The Mobile Bearer Control Function (MBCF).*

An important requirement when developing this model was that it should be independent of the radio access technology. Radio resource control and radio transport functions covering the control of the radio resources and the transport of control and user information (emission and reception) across the radio interface have been identified in separate models. The exact nature of these models and the relationship between these models and the generic functional model have yet to be defined.

3.3 UMTS Radio Aspects

Spectrum Matters

The availability of sufficient radio resources is one of the fundamental prerequisites for the legal operation of any radio system. In this respect, the World Administrative Radio Conference in 1992 (WARC'92) represented a major break-through in the work on third generation mobile telecommunications systems. WARC'92 designated the following 230 MHz of spectrum for use by FPLMTS:

1885 – 2025 MHz (140 MHz)

2110 – 2200 MHz (90 MHz)

Within these bands, two 30 MHz bands have been identified for use in Europe by the UMTS satellite component (if implemented), namely 1980 – 2010 MHz (earth to space), and 2170 – 2200 MHz (space to earth), see Figure 5.

It is worth noting that these frequency bands are not exclusively reserved for UMTS or FPLMTS. In some geographical areas UMTS will be limited to operate in parts of the bands, and in some areas UMTS will be required to share spectrum with already existing services (e.g. fixed microwave links). Moreover, the satellite component may be implemented in the satellite bands only, whereas the terrestrial component may be implemented in both the terrestrial and satellite frequency bands.

In principle, parts of the bands indicated in Figure 5 should be made available in the year 2000. However, the exact frequency bands that UMTS will use in various geographical areas, and the time they are to be made available, are still to be defined, and are subject to current CEPT studies. These studies are planned to be concluded in 1995.

Radio interface design methodology

The following procedure has been adopted for the design of the UMTS radio interface:

1 A set of requirements for the radio interface is defined, based on system and services requirements for UMTS [11].

2 Using the radio interface requirements, a selection procedure for the choice of radio access principles (multiple access technique, transmission bandwidths, duplex technique, modulation method, etc.) is developed and verified against existing systems [12].

3 Proponents of candidate radio access technologies are asked to submit their proposals, which will be passed through the selection process and given a relative ranking (the process has been dubbed "The World Championship in Radio Access Technology").

4 Using the relative ranking, a set of candidates are selected and used when making the radio interface specifications.

The radio requirements, approved by SMG5 in December 1994, are the only part of this procedure which so far has been completed. Step 2 is expected to be completed in mid 1996 and step 3 towards the end of 1996.

Requirements on the UMTS radio interface

Key aspects of the radio interface requirements are the identification of radio operating environments, characterised by radio propagation characteristics, relative speed between the terminal and the radio

Figure 6 A possible GSM migration path towards UMTS

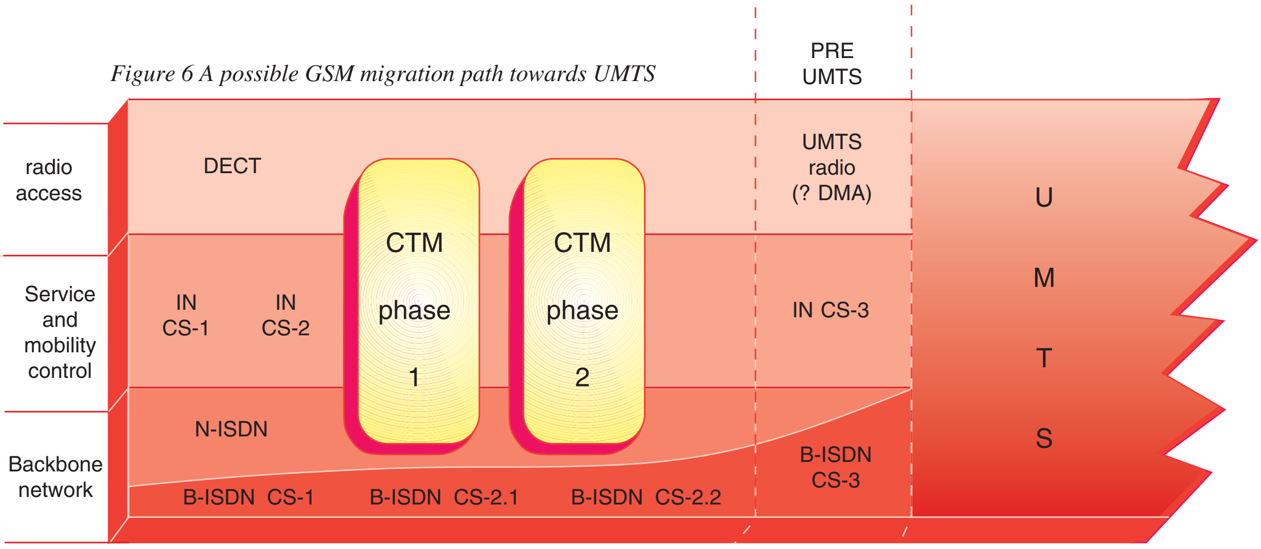
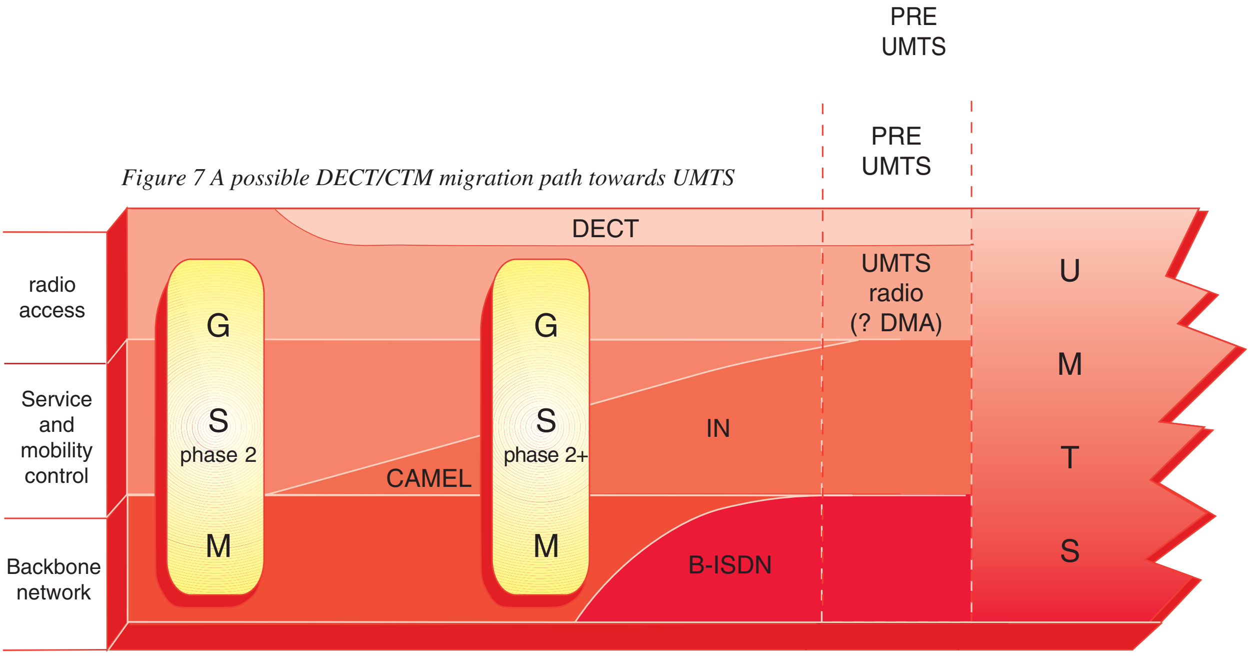


Figure 7 A possible DECT/CTM migration path towards UMTS



port and user traffic conditions, and the identification of the UMTS bearer service classes supported in the different radio operating environments. 14 radio operating environments have been identified, including business, neighbourhood and home indoor environments, urban vehicular, urban pedestrian and rural outdoor environments, a local high bitrate environment and several satellite indoor and outdoor environments. The highest user bitrates (1920 kbit/s) will be offered using Class A, B or D bearer services in the business indoor, home and local high bitrate environments, the lowest user bitrates (1.2 kbit/s) using Class A, B, C or D bearer services in the satellite indoor environment.

The choice of radio access technology

The choice of radio access technologies defines to a large extent the performance of the radio sub-system. It is also a topic in which the manufacturing industry and the operators have already made considerable investments and has therefore been subject to considerable controversy and much interest world-wide.

At the core of the selection procedure is the evaluation of the relative merits of the three multiple access technologies *frequency division multiple access (FDMA)*, *time division multiple access (TDMA)* and *code division multiple access (CDMA)*. The initial design objective for the radio interface was to select one of these technologies offering the optimum compromise of performance in all radio operating environments, in order to fully exploit the allocation of a world-wide frequency band. However, it is now commonly accepted that a set of radio access technologies may be required in order to meet the UMTS radio requirements. UMTS may therefore be expected to support a family of radio interfaces tailored to different radio operating environments. Continuos UMTS access across radio operating environments will require support from the emerging technologies of *multi-mode terminals* and *software radios*. To facilitate these new technologies, considerable emphasis is now placed on maximising the commonality between the potential radio interfaces rather than designing one common radio interface.

4 Migration aspects

UMTS will be introduced into a telecommunications market with a number of existing and planned systems of varying degrees of maturity. In order to reduce the cost of introducing UMTS in this market and to ease the transition of the customer base of existing mobile systems to UMTS, it will be advantageous to reuse parts of the already existing physical infrastructure. This suggests that a phased introduction of UMTS will be desirable, where the initial phases of UMTS are introduced as limited enhancements to the existing systems, which in later phases are evolved towards full UMTS support. The description of the ways in which an existing system can evolve towards full UMTS support are known as *migration paths*.

The two main candidates for migration towards UMTS in Europe are GSM/DCS 1800/DCS 1900 and DECT/ CTM, both systems operating close to the frequency band identified for UMTS. Figures 6 and 7 outline possible migration paths for GSM and DECT/CTM, focusing on evolution of the three system components: radio access technology, service and mobility control, and access and backbone network.

The GSM migration path has the following main steps:

- *New radio access technology* will be needed to fulfil the requirements on the UMTS radio interface (e.g. bitrates up to 2 Mb/s).
- *Multi-mode GSM terminals,* extending GSM coverage to the private and residential service environments, will be required. The already announced introduction of dual-mode DECT/GSM terminals is an important step on this path.
- *GSM/IN integration* to support a more flexible mechanism for service creation and control is required. A first step on this path has already been taken through the introduction of the CAMEL feature (Customised Application for Mobile network Enhanced Logic) in GSM phase 2+, in order to introduce IN services to roaming GSM users.
- *Support for B-ISDN* in the access network.

For the DECT/CTM migration path, the following main steps can be identified:

- *New radio access technology*.
- *Multi-mode terminals,* extending DECT coverage to the public service environment.
- *Evolution of the IN platform* from Capability Set 2 (CS-2) to CS-3. CTM phase 1 and CTM phase 2 can be considered as a part of this IN evolution, because the functionality of CTM is mainly dependent on IN. However, CTM also includes the radio access and backbone network, and Figure 7 shows CTM as a concept on top of IN, ISDN and the radio access.
- *Evolution from support of ISDN to support of B-ISDN*. B-ISDN will most likely evolve towards a common transport platform supporting both narrowband and broadband bearer capabilities.

It should be noted that the DECT/CTM migration path is heavily based on current fixed network technology. DECT/ CTM migration therefore provides evolution for the fixed networks towards UMTS, thus offering a means for the fixed network operators to enter the expanding mobile communications market after the expected re-regulation in 1998.

5 Conclusions

The process of designing UMTS is a lengthy one, as could be expected from the still ongoing GSM standardisation work. Technological barriers will have to be overcome, new alliances sought and numerous compromises reached before UMTS will become a reality. However, if the standardisation effort is successful, a new system fulfilling most telecommunication needs in the global mobile telecommunications market place of tomorrow looks set to emerge. Already the contours of a global third generation system with the potential to change communications patterns all over the world can be seen.

6 References

1 WARC'92. *The world administrative radio conference for dealing with frequency allocations in certain parts of the spectrum.* Malaga-Torremolinos, 1992.

2 Svebak, O D, Bjørnland, D F. *The universal mobile telecommunication system (UMTS) : status and quo vadis.* Kjeller, Telenor Research, 1995. (Report R 18/95.)

3 ETSI DTR/SMG-050101. *UMTS objectives and overview, version 2.1.0.* Sophia Antipolis, 1995.

4 ETSI DTR/SMG-050103. *UMTS system requirements, version 2.0.0.* Sophia Antipolis, 1994.

5 ETSI DTR/SMG-050201. *Framework for services to be supported by UMTS, version 2.0.0.* Sophia Antipolis, 1994.

6 ETSI DTR/SMG-050501. *Objectives and framework for the telecommunication management network (TMN), version 2.0.2.* Sophia Antipolis, 1995.

7 ETSI DTR/SMG-050301. *Framework of network requirements, interworking and integration for UMTS, version 1.3.0.* Sophia Antipolis, 1995.

8 Race common functional specifications D733. *Network aspects, issue D1.* Brussels, 1994.

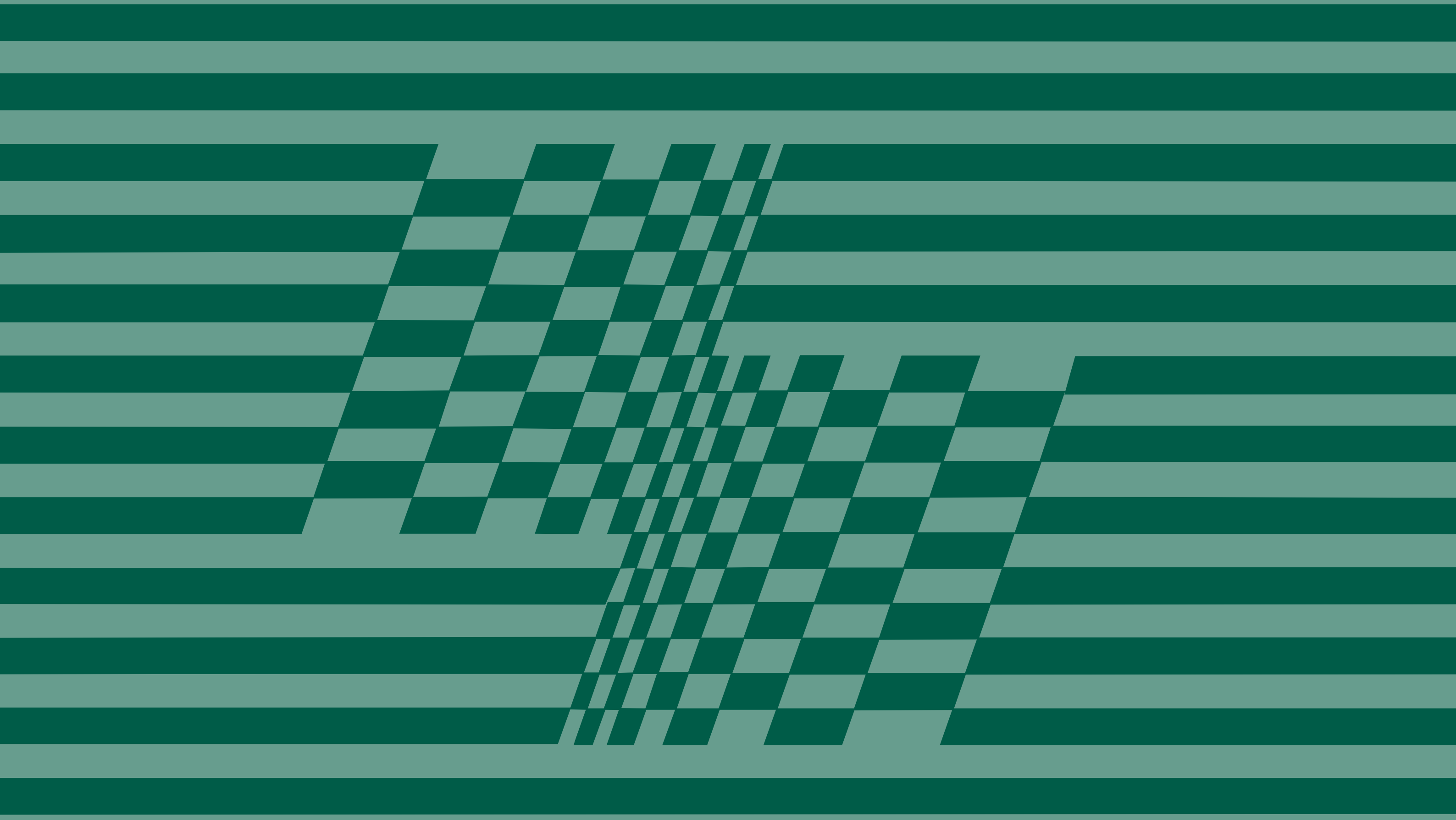
9 ITU-T. Recommendation I.130. *Method for characterisation of Telecommunication services supported by an ISDN and network capabilities of an ISDN.* Melbourne, 1988.

10 CCITT Recommendation Q.1214. *Distributed functional plane for intelligent network CS-1.* Geneva, 1992.

11 ETSI DTR/SMG-050401. *Overall requirements on the radio interface(s) of the UMTS, version 2.0.0.* Sophia Antipolis, 1995.

12 ETSI DTR/SMG-050402. *Selection procedures for the choice of radio transmission technologies for UMTS, version 0.8.2.* Sophia Antipolis, 1995.

Special



Mathematical model and algorithms used in the access network planning tool FABONETT

BY RALPH LORENTZEN

1 Introduction

The PC-based service access point network structure planning tool FABONETT has been created through a co-operative effort between Telenor Research (TR), the Planning Division in Telenor Region Oslo, and Det Norske Veritas Industry (DNVI). FABONETT has been developed for the regional planners who are responsible for access network planning in Telenor.

FABONETT is an *integer programming optimisation model* which attempts to find the most economical service access point structure which meets all capacity and connectivity requirements. The integer program is solved using a combination of a branch and cut algorithm with dynamic variable and constraint generation and heuristics.

The planner can choose which categories of variables FABONETT will treat as integer variables in the branch and cut phase of the algorithm and which variables FABONETT will determine by using heuristics. The user interface allows the planner to make modifications to the solution found by FABONETT and check feasibility and cost.

FABONETT has a graphical and table based user interface, and most of the development work has been directed to the implementation of this interface. The user interface is described in [1]. In this paper, however, only the mathematical model and algorithms used in FABONETT are described.

2 The service access point structure design problem

Here, a short description will be given of the service access point structure design problem FABONETT attempts to solve.

We are given a *local switch (LS)* and a set of *main distribution points (MDs)* together with a set of what we call *fibre subscribers (FSs)* that must be connected to the local switch with fibre cables. The MDs and the FSs can be connected directly to *LS* or via *service access points (SAPs)* where multiplexing is done.

FABONETT operates with cable and radio connections while the optimisation is based on cable connections only. The planner can, however, use FABONETT to evaluate wholly specified hybrid radio/cable network designs. Since this paper describes the optimisation model only, it will be assumed that all connections are made with cables.

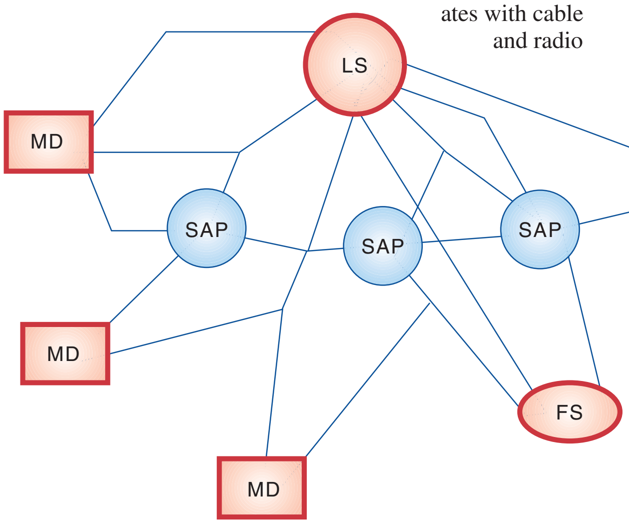


Figure 1 Input network structure

The cables can be *copper cables* or *fibre cables*. The cables are following *paths.* A path consists of *sections* which are characterised by their *cost, length, section type* and one or more *section codes*. Typical section types are *conduits* and *ducts* (existing or new), *trenches* of different types and *air cable sections.*

A section code is simply a positive integer. Two sections share a section code if events causing damage to the two sections are assumed to be positively correlated. Paths inherit section codes from the sections they consist of. When two independent paths between a SAP and *LS* must be established, they cannot share the same section code.

FABONETT operates with three different SAP types:

- A *SAP1* is connected to *LS* with one sequence of single fibre pairs.
- A *SAP2* is connected to *LS* with two sequences of single fibre pairs following two distinct paths which do not have any section code in common.
- A *SAPR* is connected to *LS* with one sequence of copper cables following one single path.

A *SAPF* denotes either a SAP1 or a SAP2.

If an MD is connected directly to *LS,* then the connection is made with copper cables. If an MD is connected to *LS* via a SAP, then the connection to the SAP is made with copper cables. An MD is characterised by the subscribers connected to it. Some subscribers require that if their MD is connected to a SAP, it has to be a SAP2.

A circuit connecting an FS to *LS* belongs to one of two types, namely *regular circuits* and *singular circuits*. The regular circuits can be connected directly to *LS* or indirectly to *LS* via a SAPF. The regular circuits which are connected to *LS* via a SAP will be multiplexed together with other circuits at the SAP. Some FSs require that if their regular circuits are connected to *LS* via a SAP, it has to be a SAP2. Singular circuits are connected directly to *LS.*

Based on the location of *LS,* locations of MDs and FSs to be connected to *LS,* locations of existing cables, locations of existing and candidate sections and locations of existing and candidate SAPs, FABONETT tries to find the least costly network design.

FABONETT formulates the design problem as an integer program which is solved by a combination of linear programming, row and column generation and fixation of variables.

In Figure 1 is shown an input network structure with candidate SAPS and sections.

In Figure 2 the same structure is shown together with chosen SAPs and connections.

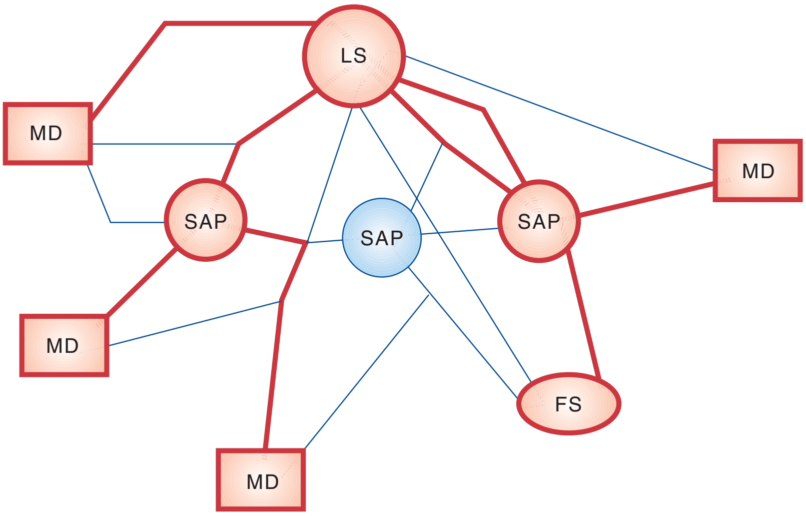


Figure 2 Input network structure with chosen design

FABONETT does not invent possible locations for candidate SAPs and candidate path sections. All candidate SAPS and candidate sections must be provided by the user.

Since FABONETT does not necessarily solve the SAP structure design problem to a theoretical optimum, the planner must inspect the solution and sometimes make FABONETT reruns with slightly altered input. The planner may for example question FABONETT's selection of a particular SAP candidate and wish to make a rerun with this SAP excluded. FABONETT's input format makes this possible without erasing the SAP candidate from the input. Or the planner may question the correctness of connecting a particular FS directly to *LS.* A rerun may then be made where the planner specifies which SAPs this FS should be allowed to connect to.

The planner has a problem of a dynamical nature. In establishing the best network structure he must take the development in demand structure over time into consideration.

FABONETT is a static 'one shot' model. Some simple features for 'dynamical use' are, however, built into FABONETT. The planner can give selected SAP candidates and path sections the label 'preferred' and give them a bonus. It is recommended to make two FABONETT runs. First a 'future run' should be made where the traffic input represents some future point in time. Then a main run is made where some or all the SAP candidates and path sections which were chosen in the future run are labelled 'preferred' and given a suitable bonus.

3 FABONETT input in broad terms

The complete input to FABONETT in the form that the planner has to present it in tables and graphs on background maps is described in detail in [1]. Still, in order to give an impression of what input data are needed to run FABONETT, an overview of main input data items is given below. Not all of these data items are used in the integer programming model, and it will not be apparent how the individual data elements relate to the model. In the next section, however, a detailed description will be given of the input which is directly used to formulate the integer program.

- *LS:*
  - location
  - cost per line directly from MDs
- Candidate SAPs:
  - location
  - status ('existing', 'free', 'must be established', 'excluded', 'preferred')
  - type
  - bonus (if status is 'preferred')
- MDs:
  - location
  - SAP type requirement
  - number of subscribers of various types
  - number of 2 Mb/s circuits coming from subscribers
  - distance to furthest business subscriber (restricts the distance from MD to SAP)
- Fibre subscribers
  - location
  - circuit requirements (# 2 Mb/s for regular and # fibre pairs for singular circuits)
  - which candidate SAPs the subscriber can be connected to
- Nodes (end points of sections which are not locations of LS, candidate SAPs, MDs or FSs):
  - location
  - cable splicing ('possible', 'not possible')
  - signal regeneration ('possible', 'exists', 'not possible')
- Sections:
  - location
  - length
  - type
  - status ('existing', 'free', 'must be established', 'excluded', 'preferred')
  - bonus (if status is 'preferred')
  - section codes
- Cables:
  - type
  - number of available pairs
  - sections the cable is placed in
- SAP types:
  - capacity towards MDs (maximum number of lines)
  - whether it is a SAP1, SAP2 or SAPR
  - yearly O&M cost
  - sundry costs for transmission equipment, cabinet, power, connections and RSU/RSS
- Subscriber types:
  - traffic measured in erlangs
- Cable types:
  - medium (copper or fibre)
  - number of pairs
  - cost per metre
  - economic life time
  - in what section types this cable type can be used
- Section types:
  - cost per metre.

4 The integer programming model

Here we are going to describe the core of FABONETT, namely the integer programming model.

We will first present the part of the input to FABONETT which is used to form the constraints in the model. Then the variables and the constraints will be described and the form of the cost function will be outlined.

For the constants needed to formulate the constraints, the following notation is used:

r(F) denotes the number of fibre pairs carrying regular circuits to FS *F* (normally 0 or 1).

s(F) denotes the number of fibre pairs carrying singular circuits to FS *F*.

P(M) denotes the required number of copper pairs connected to MD *M*.

TO(M) denotes the number of 2 Mb/s circuits connected to MD *M*.

k_{2S} denotes the number of copper pairs required per 2 Mb/s circuit between a SAPR and *LS*.

$f(S) = 2$ if SAP *S* is a SAP2, and = 1 otherwise.

$f(F) = 2$ if the only SAP type FS *F* can be connected to is SAP2, and = 1 otherwise.

p_{fij} denotes the maximum number of new fibre pairs that will be considered for section *(i,j)*.

p_{cij} denotes the maximum number of new copper pairs that will be considered for section *(i,j)*.

f_{ij} denotes the number of available pairs in existing fibre cables in section *(i,j)*.

k_{ij} denotes the number of available pairs in existing copper cables in section *(i,j)*.

TRAF(M) denotes the traffic measured in erlangs in the lines from MD *M*.

E_0 and E denote the constant term and rate of increase respectively in the linearised erlang function which expresses the number of channels needed as a function of traffic with a given blocking probability. If this probability is p, then $E_0 = (1 - p)/p$ and $E = (1 - p)$.

$\mathtt{M}\,(S)$ denotes a minimal set of MDs which can be connected to SAP *S* and which together exceed the capacity of SAP *S*.

If a section (usually duct sections) is not given a section code, the optimisation module in FABONETT duplicates the section and assigns distinct section codes to the original and the duplicate sections. This is done in order to secure that two independent circuits connecting *LS* to a SAP2 do not share the same fibre cable.

The problem variables will now be described:

$x_S = 1$ if SAP *S* is established and = 0 otherwise

$x_{MS} = 1$ if MD *M* is connected to *LS*/SAP *S* and = 0 otherwise

$x_{FS} = 1$ if the regular circuits of FS *F* is connected to *LS*/SAP *S* and = 0 otherwise

x_{MGp} is the number of copper pairs to MD *M* from *LS*/SAP node *G* via path *p*

x_{FGp} is the number of fibre pairs to FS *F* from *LS*/SAP node *G* via path *p*

x_{fGp} is the number of fibre pairs from *LS* to SAPFs in node *G* via path *p*

x_{cGp} is the number of copper pairs from *LS* to SAPRs in node *G* via path *p*

$x_{ij} = 1$ if new cables are placed in section *(i,j)* is and = 0 otherwise (for sections with cost ≠ 0 only)

x_{fij} is the number of new fibre pairs placed in section *(i,j)*

x_{cij} is the number of new copper pairs placed in section *(i,j)*

y_M, y_F, y_{fij}, y_{cij} and y_{Gs} are slack variables with a high cost.

All variables are required to be integer. The individual cable types are not modelled explicitly. The detailed cabling is determined in a postprocessor.

The problem constraints are formulated below:

(1) $\Sigma p\, x_{MGp} - P(M)\, \Sigma_{S \in G}\, x_{MS} = 0$

(2) $\Sigma_p\, x_{FGp} - (r(F) + s(F))\, \Sigma_{S \in G}\, x_{FS} = 0$ where the term *s(F)* is included only if $G = LS$

(3) $\Sigma_S\, x_{MS} + y_M = 1$

(4) $\Sigma_S\, x_{FS} + y_F = 1$

(5) $x_S - x_{MS} \geq 0$

(6) $x_S - x_{FS} \geq 0$

(7) $(|\mathtt{M}\,(S)| - 1)\, x_S - \Sigma_{M \in \mathtt{M}\,(S)}\, x_{MS} \geq 0$ for all minimal $\mathtt{M}\,(S)$s

(8) $\Sigma_v\, x_{fGp} - \Sigma_{S \in G \cap SAPF}\, f(S)\, x_S = 0$

(9) $\Sigma_p\, x_{cGp} - k_{2S}\, \Sigma_{S \in G \cap SAPR}\, [E_0 x_S / 30 +$
$\Sigma_M\, (E \cdot TRAF(M) / 30 + TO(M))\, x_{MS}] \geq 0$

(10) $-\Sigma_{G, \{p;(i,j) \in p\}}\, x_{fGp} - \Sigma_{F,G, \{p;(i,j) \in p\}}\, x_{FGp} + f_{ij} + x_{fij} + y_{fij} \geq 0$

(11) $-\Sigma_{G, \{p;(i,j) \in p\}}\, x_{cGp} - \Sigma_{M,G, \{p;(i,j) \in p\}}\, x_{MGp} + k_{ij} + x_{cij} + y_{cij} \geq 0$

(12) $-\Sigma_{\{p; s \in p\}}\, x_{fGp} + \Sigma_{S \in G \cap SAPF}\, x_S + y_{Gs} \geq 0$ for section code *s*

(13) $p_{fij}\, x_{ij} - x_{fij} \geq 0$

(14) $p_{cij}\, x_{ij} - x_{cij} \geq 0$

(15) $x_{ij} - x_{ij}^{\,d} \geq 0$ where $x_{ij}^{\,d}$ represents the duplicate section of (i,j)

The constraints (1) express that all copper pairs from MD *M* to SAPs (or to *LS*) at node *G* must follow a path from *M* to *G*.

The constraints (2) express that all fibres from FS *F* to SAPs (or to *LS*) at node *G* must follow a path from *S* to *G*.

The constraints (3) express that MD *M* must be connected to a SAP (or to *LS*). The slack variables y_M with high costs are introduced in order to avoid infeasibilities when solving the integer program.

The constraints (4) express that FS *S* must be connected to a SAP (or to *LS*). The slack variables y_F with high costs are introduced in order to avoid infeasibilities when solving the integer program.

The constraints (5) express that if MD *M* is connected to SAP *S*, then SAP *S* must be established.

The constraints (6) express that if FS *F* is connected to SAP *S*, then SAP *S* must be established.

The constraints (7) express that the number of subscriber lines connected to SAP *S* must not exceed *S*'s capacity.

The constraints (8) express that there must be a sufficient number of fibre pairs following paths from node *G* to *LS*.

The constraints (9) express that if a SAPR is established in node G, the number of copper pairs from G to *LS* must be sufficiently large to serve the traffic with a blocking probability less than or equal to a prescribed upper limit.

The constraints (10) express that there must be sufficient number of fibre pairs in section *(i,j)* to support all fibre requirements in it. Slack variables y_{fij} with high costs are introduced in order to avoid infeasibilities when solving the integer program.

The constraints (11) express that there must be sufficient number of copper pairs in section *(i,j)* to support all copper requirements in it. Again slack variables y_{cij} with high costs are introduced in order to avoid infeasibilities when solving the integer program.

The constraints (12) express that at most 50 % of the fibre pairs from a SAP2 to *LS* can pass through a section with section code *s*. Slack variables y_{Gs} with high costs are introduced in order to avoid infeasibilities when solving the integer program.

The constraints (13) and (14) express the obvious fact that if a new (fibre or copper) cable is put in a section, the section will contain new cables.

The constraints (15) express that new cables can be placed in a duplicate section only if new cables are put in the corresponding original section.

The form of the cost function to be minimised is simple in principle:

$$(18)\ \Sigma c_S x_S + \Sigma c_{MS} x_{MS} + \Sigma c_{MGp} x_{MGp} + \Sigma c_{FS} x_{FS} + \Sigma c_{SGp} x_{SGp} \\ + \Sigma c_{fGp} x_{fGp} + \Sigma c_{cGp} x_{cGp} + \Sigma c_{ij} x_{ij} + \Sigma c_{fij} x_{fij} + \Sigma c_{cij} x_{cij} + \\ M\ \Sigma(y_M + y_F + \Sigma y_{fij} + \Sigma y_{cij} + y_{Gs}).$$

Here *M* is a large number. The coefficients of the *x*-variables depend, however, on a rather large set of detailed technical and financial data which will not be dealt with here.

The solution method is a combination of:

- linear programming with dynamic path generation
- branch and cut
- cost adjustment heuristics (optional)
- variable fixing heuristics.

In the branch and cut phase x_S, x_{MS}, x_{FS} and, optionally, x_{ij} are required to be integer. For large problems the branch and cut phase may take too long time to finish if the x_{ij} variables are required to be integer. Therefore, the planner has the option of not requiring these variables to be integer in the branch and cut phase. Their values will then, together with the other variables not required to be integer in the branch and cut phase, be determined by heuristics.

5 Linear programming with dynamic path generation

The integer program is first solved as a linear program. The number of path variables x_{MGv}, x_{SGv}, x_{fGv} and x_{kGv} is so large that it is unrealistic to include them all explicitly in the model from the beginning. The relevant path variables are therefore generated dynamically during the solution process using classical column generation techniques. The following notation is used for the shadow prices associated with the constraints which involve path variables:

Constraint number:	Shadow price:
(1)	π_{MG}
(2)	π_{FG}
(8)	π_{fG}
(9)	π_{cG}
(10)	π_{fij}
(11)	π_{cij}
(12)	π_{sG}

The reduced costs for the path variables then become:

Path variable:	Reduced cost:
x_{MGp}	$c_{MGp} - \pi_{MG} + \Sigma_{ij}\pi_{cij}$
x_{SGp}	$c_{SGp} - \pi_{FG} + \Sigma_{ij}\pi_{fij}$
x_{cGp}	$c_{cGp} - \pi_{cG} + \Sigma_{ij}\pi_{cij}$
x_{fGp}	$c_{fGp} - \pi_{fG} + \Sigma_{ij}\pi_{fij} + \Sigma_s\pi_{sG}$

Since the original cost coefficients of the path variables are sums of cost components associated with each of the sections which make up the paths, the path variable of each type with the smallest reduced cost can be found by solving a shortest path problem where the length of a path section *(i,j)* is the sum of the section's cost component, π_{cij} or π_{fij}, and the π_{sG} variables relevant for the section. The shortest path problems from *M*s to *G*s, from *S*s to *G*s, and from *G*s to *LS* are all solved using Dijkstra's classical shortest path algorithm.

6 Branch and cut

In theory the integer program could be solved to an exact optimum using the branch and bound algorithm with column generation at every node in the branch and bound tree. The large number of variables and the weakness of the constraints (13) and (14) make this unrealistic, so we limit ourselves to using the branch and bound algorithm (with column generation at the nodes of the branching tree) where only a subset of the variables are required to be integer. This subset can be specified by the user, but normally the variable types required to be integer will be x_S, x_{MS}, x_{FS} and, sometimes, x_{ij}.

The constraints (13) and (14) are weak. Therefore, if we do not introduce some form of cuts, the solution time can become prohibitively long if we require the path section variables x_{ij} to be integer in the branch and bound process. Alternatively, If we do not require them to be integer, the heuristics may lead to suboptimal solutions where trench section costs have been underestimated compared to cable costs. In order to partially remedy this we heuristically add constraints (cuts) at the nodes in the branch and bound tree.

All trench sections containing cables are duplicated and the two duplicates will receive the same section code. One of the two duplicates contains no cables and is denoted as *unestablished.* The other duplicate is denoted as *established* and contains all the existing cables, but no new cables can be placed in this duplicate.

We introduce the variables z_M where z_M =1 if no path connecting *M* to *LS/SAP* run in unestablished trenches and = 0 otherwise, and the constraints

(19) $\Sigma_{G,p \text{ not in unestablished trenches}}\, x_{MGp} - P(M)\, z_M \geq 0$

with corresponding shadow prices π_M.

At a branch and bound tree node we then consider the network of path sections and give capacity x_{ij} to unestablished trench sections *(i,j)* and capacity *P(M)* to the other sections.

For each *M* we augment the network with edges with capacity *P(M)* between LS/SAP candidates and a supernode *SN* (see Figure 3).

Then we find maximum flow between *M* and *SN* with a min cut *C* and check if

(20) $z_M + \Sigma_{\text{unestablished trenches } (i,j) \text{ in } C}\, x_{ij} \geq 1$

If not, this inequality is added to the LP which is reoptimized (with path generation) in the branch and bound tree node.

If $\pi_M > 0$, and the shortest path traverses at least one unestablished trench section, then we must run an additional shortest path generation where we do not use unestablished trenches. Such a path receives a bonus π_M.

Type (20) cuts may erroneously prevent digging of unestablished trenches. This can happen when it is possible to connect an M to *LS/SAP* without using unestablished trenches.

7 Cost adjustment heuristic

In the case where the x_{ij} variables are not required to be integer in the branch and cut phase a cost adjustment heuristic is used. The linear program (with fixed x_S, x_{MS} and x_{FS} variables) is run once more before variable fixing is initiated. In this run the cost coefficients of the x_{ij} variables are divided by the fractional value they had in the previous run. x_{ij} variables with cost 0 are not affected, and variables with positive cost which got the value 0 in the previous run will be given a cost equal to infinity and are in effect removed from the problem.

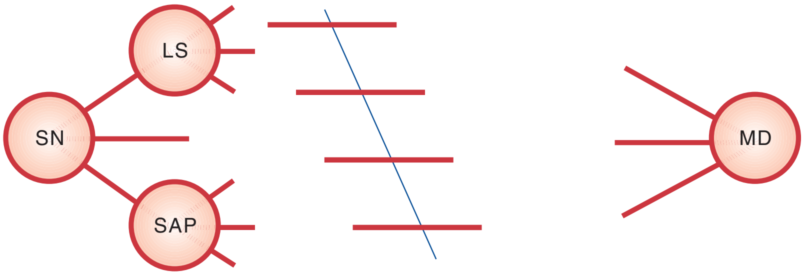


Figure 3 Network at branch and bound tree node

8 Variable fixing heuristics

At the end of the branch and bound phase the variables of type x_S, x_{MS}, x_{FS} (and, optionally, x_{ij}) are integer. They are then fixed at their current values. At the start of the variable fixing phase, if the variables of type x_{ij} are not integer, they are forced to integer values. This is done by selecting the x_{ij} with the highest fractional value, fixing this variable at its ceiling, and solving the linear program (with column generation) again. This process is continued until all the x_{ij} variables are integer. Then the x_{MGv}, x_{SGv}, x_{fGv} and x_{kGv} variables are forced to integer values. This is done by selecting the variable of one of these types with the highest fractional value, fixing this variable at its ceiling, and solving the linear program (without column generation) again. This process is continued until all the x_{MGv}, x_{SGv}, x_{fGv} and x_{kGv} variables are integer.

A stage is now reached where the integer program is solved in the sense that all SAPs, path sections and cables to be used are selected, and it is decided which MDs and FSs should be connected to which SAPs (or *LS*). It remains only to use the cables to make all the necessary connections. This is not explicitly modelled in the integer program, and is therefore done in a post-processing of the solution.

9 Typical problem size and solution time

A medium size problem will have 10,000 – 15,000 constraints and 4,000 – 5,000 initial variables. The path variable generation process will typically generate around 2,000 path variables. The solution time for such a problem will be 10 – 20 minutes on a 486 33 MHz PC.

Acknowledgement

The author wishes to thank Harald Moseby who has contributed many ideas during the development of the algorithms and who has programmed the algorithms for PC.

Reference

1 Moseby, H, Kirkeberg, G. *FABONETT user's guide.* Høvik, Det Norske Veritas Industry, 1993. (Report 93-3670.) (In Norwegian)

Performance of packet switched services in spread-spectrum radio networks

BY TORE J BERG AND PEDER J EMSTAD

The throughput and delay of packet switched connectionless services in a fully connected single-channel spread-spectrum radio network operating under a random access scheme is considered. Firstly, a layered model and its layer services and protocols are presented. The analytical models for performance studies include the most important parameters of a spread-spectrum packet radio; the turn time, the synchronisation delay and the capture effect and some protocol time parameters. Throughput at maximum load is found and delay is studied under some independence assumptions. The models are validated by simulations.

1 Introduction

This study addresses the steady-state performance of packet switched services in a single-channel spread-spectrum radio [1] network using a random access protocol. Most of the studies in the literature on various models of packet radio networks assume that each node can hold just one packet ready for transmission. Pioneering works under such assumptions were done in [3] and [4]. In [5] a CSMA network was studied and under several assumptions a product form solution was obtained. Throughput/delay analysis of packet radio networks with infinite buffer size do not generally yield to product form solutions. However, the study of a two node network in [6] gives a product form solution under some constraints, but becomes quite complicated for a larger network. Generally, approximative methods have to be used. One such study is reported in [7] where each node is modelled as an M/G/1/∞ queue with additional service to the first customer in the idle period. The model is solved by iteration.

The theoretical model developed in this study also resorts to approximations and our aim is that the model shall give good accuracy around the operating point of a real packet radio network. The achieved accuracy is checked by simulations. The nodes in the network are modelled as independent M/G/1/∞ queues with set-up times [8] following the ideas in [7]. The fact that the radio channel is shared by the nodes is incorporated through the packet service time which then will be a function of the traffic level at the other nodes.

In a real system the propagation loss is a function of the distance between the transmitter and the receiver. We model the received power to distance relationship as $1/r^{\beta}$, where β is a constant depending on the terrain. The spatial distribution model used in the validation of the theoretical model is a regular grid network.

A fundamental assumption taken in developing the theoretical model is that all the nodes behave identically, and therefore, we only cater for uniformly distributed traffic pattern.

The term system parameters is used to address values concerning radio properties, protocol parameters and network size. Offered user traffic is characterised by the packet arrival and the packet length. Packet arrivals are modelled by a Poisson process.

Chapter 2 presents a layered model of the network in conformance with [10], and describes its layer services and protocols.

Chapter 3 forms a throughput model that incorporates the radio parameters that have most impact on network performance. The chapter is ended by a case study.

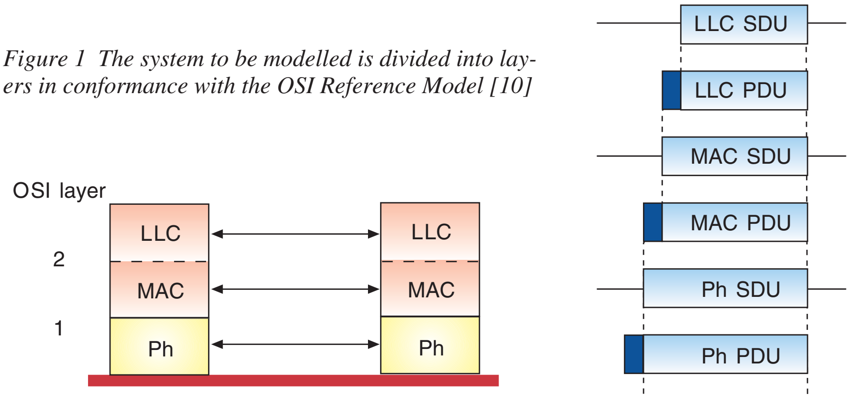
The network transit delay is mainly the sum of two stochastic components; the time taken to get access to the radio channel and send a packet, and the time a packet must be queued before it is taken under service. Chapter 4 considers the packet service time and the time spent waiting for service.

2 A layered model

The description of the system to be modelled is based on the layered model in Figure 1. The physical layer protocol is given by the radio and is specified in section 2.1. This layer provides communication services to the next upper layer, the link layer. The link layer is divided into two sublayers; the Medium Access Control (MAC) and the Logical Link Control (LLC). The objective of the MAC protocol is to give the nodes access to the radio channel in a fair manner. The LLC layer performs queuing.

In the context of Figure 1, the user traffic arrives by LLC-UNITDATA requests [11] at the LLC layer and will, if successfully transferred by the system, arrive at the addressed destination as an LLC-UNITDATA indication. To facilitate the transfer of LLC Service Data Units (SDUs) the LLC layer entity uses the service of the MAC layer; the MAC-UNITDATA service primitive.

Figure 1 The system to be modelled is divided into layers in conformance with the OSI Reference Model [10]



Each layer protocol needs to add protocol control information (PCI) to perform their functions. Therefore, the user packet length, or more precisely the LLC SDU, will have grown in size when finally transferred on the radio channel as a Physical layer Protocol Data Unit (Ph PDU).

The use of this model allows us to give an unambiguous definition of the performance measures used. We now turn to the specification of the individual layer protocols.

2.1 The physical layer

The first opportunity for a receiving radio to detect a transmission is after the time delay for the transmitting node to turn from an idle state to transmit mode (t_{turn})

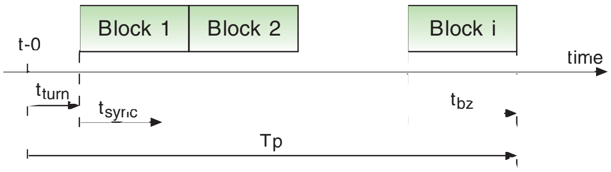


Figure 2 A radio switches to transmit mode at t = 0 and t_v time units later its neighbours detect a busy channel

plus the length of the preamble *(t_{sync})*, see Figure 2. Within this period of time no receivers are able to detect a transmission. This period, $t_{turn} + t_{sync}$, is defined as the vulnerable period t_v.

Practical use of radio demands implementation of a Forward Error Correction (FEC) coder to reduce the channel bit error rate to a level suitable for data transmissions. We assume use of a block coder of size t_{bz}. For simplicity, protocol control information (e.g. fields needed for error detection) is neglected in this study. We define the channel packet length as

$$T_P \triangleq t_{turn} + N_P \cdot t_{bz} \qquad (1)$$

where N_P is the random number of blocks in the user packets.

Let B_k be the probability that the first transmission succeeds given that k nodes transmit simultaneously. We assume that none of the nodes are influenced by background noise or local noise leading to the two trivial cases $B_1 = 1$ and $B_n = 0$ when n is the number of nodes in the network. If a node B is locked on a transmission from a node A, all other overlapping transmissions will be seen as noise from node B's point of view. Let $SNR_{NTx,A \to B}$ be the signal to noise ratio at node B when B is *locked* on a transmission from node A. Then

$$SNR_{NTx,A\to B} = 10\log\left(\frac{1}{r^{\beta}_{A,B}\sum_{\forall j\in NTx} r^{-\beta}_{B,j}}\right)\text{[dB]} \qquad (2)$$

where NTx is the set of interfering sources and $r_{i,j}$ is the distance between node i and node j.

Let $SNR_{capture}$ denote the resistance to any interfering signals. A radio is able to keep receiving an already captured packet until the total noise of the interfering signals reaches this threshold, see Figure 3. The solid line shows the idealised curve while the dotted line shows the curve for a typical radio. The curve depends on the modulation and coding of the radio used. The justification of this curve against a radio implementation is outside the scope of this study.

A multi block packet has a higher $SNR_{capture}$ than a single block packet but we will use the same $SNR_{capture}$ value for both.

Define the following indicator function

$$Isuccess_{NTx,A\to B} = \begin{cases} 1 \text{ if } SNR_{NTx,A\to B} \geq SNR_{capture} \\ 0 \text{ if } SNR_{NTx,A\to B} < SNR_{capture} \end{cases} \qquad (3)$$

that is, 1 is returned if node B is able to successfully demodule the signal from A while all the nodes in the set NTx are transmitting. Otherwise, 0 is returned. Define

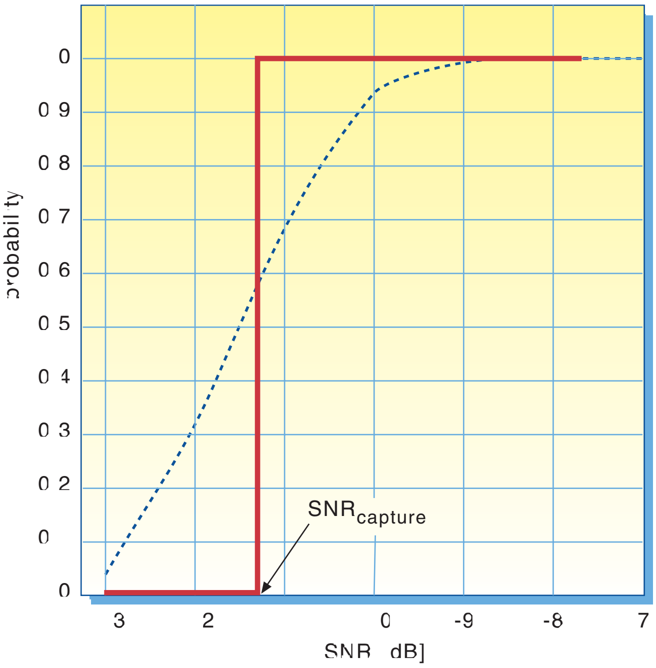


Figure 3 Probability of successful demodulation as function of the SNR level

$B_{k,A\to B} \triangleq$ the probability of successful demodulation of the transmission $A \to B$ when k additional nodes transmit and given that node A was the first node to

$$\text{transmit} \quad = \frac{1}{|S_k|}\sum_{\forall j\in S_k} E\left[Isuccess_{j,A\to B}\right]$$

where S_k is all the subsets $N_{Net} - \{A, B\}$ of size k. | | is the cardinality function and N_{net} is the set of all nodes in the network. Define

$B_{k,A} \triangleq$ the probability that one random link out from node A sustains k additional transmissions

$$= \frac{1}{n-1}\sum_{\forall i\in N_{net}-\{A\}} B_{k,A\to i}$$

$B_{k,A}$ is not identical for all nodes in a regular grid network of size greater than four. If we neglect this fact, simplified theoretical expressions are achieved without loss of generality; the throughput model may easily be extended to use different $B_{k,A}$ (the delay model does not depend on B_k). Therefore, B_k is taken as an average over all the nodes

$$B_k = \frac{1}{n}\sum_{\forall i\in N_{net}} B_{k-1,i} \quad \text{for } 2 \leq k \leq n \qquad (4)$$

To get numerical values we set β = 4 and $SNR_{capture} = -11.0$ dB. Table 1 shows some results for regular grid networks where the rows marked "exact" give values given by (4).

We have two observations. Firstly, a good approximation is to use the same B_k-values for network sizes within the range 9 to

Table 1 *The capture probability B_k for different network sizes*

network sizes (n)	Formulae	B_2	B_3	B_4	B_5
25	exact	0.8168	0.6629	0.5425	0.4581
	q^{k-1}	0.8168	0.6672	0.5449	0.4451
9	exact	0.8413	0.7249	0.6381	0.5651
	q^{k-1}	0.8413	0.7078	0.5955	0.5010

25. Secondly, setting $B_k = q^{k-1}$ is also a reasonable approximation when q is set to B2 calculated by (4). The effect of background noise may be modelled as $B_k = p_{noise}q^{k-1}$.

2.2 The MAC layer

The MAC entity can only hold one outgoing MAC PDU at a time, and in the state *idle* this buffer is empty and no adjacent transmission is detected (by the Ph layer entity and then signalled to the MAC layer entity). If the radio has detected a transmission while the MAC buffer is empty, the MAC state is *receiving*. In the state *transmitting* the node has grabbed the channel and transmits the packet under service. In state *waiting* the MAC entity has a PDU ready for transmission but further action is prevented by an ongoing adjacent transmission. A MAC entity in the *scheduling* state has drawn a random access time delay D and is awaiting its expiration. The random access delay is taken from the uniform distribution given by the probability distribution function

$$F_D(t) = \begin{array}{l} 0 \text{ for } t \leq t_l \\ 1/t_u - t_1\ (t - t_l) \text{ for } t \ ?\ [t_l,t_u] \\ 1 \text{ for } t \geq t_u \end{array} \qquad (5)$$

How to select a t_u value will be considered later.

2.3 The LLC layer

No automatic repeat request protocol is considered in this study, so the LLC layer protocol becomes very simple. One function is to perform queuing of the user data that arrives via the LLC-UNITDATA request and whenever the MAC layer is ready to serve a MAC SDU, the LLC layer entity shall immediately issue a MAC-UNITDATA request containing the first queued data unit (FIFO operation).

3 Throughput

In this chapter we assume heavy load, every node has at least one packet ready for service at any time instance. We look at the time period between time instances where deliveries are completed, t_1 in Figure 4 and define

$$Y \overset{\Delta}{=} \text{time period between successful delivery}$$

$$p_{net} \overset{\Delta}{=} \text{probability that the first data packet is correctly received}$$

$$K_{Tx} \overset{\Delta}{=} \text{the number of simultaneous transmissions during one channel access}$$

$$C_I \overset{\Delta}{=} \text{the channel idle period}$$

$$C_B \overset{\Delta}{=} \text{the channel busy period.}$$

The number of blocks delivered per time unit is $E[N_P]$ / $E[Y]$ because a packet contains N_P blocks. We get the normalised network throughput λ_{max} when dividing by the physical layer transfer rate $1/t_{bz}$

$$\lambda_{max} = t_{bz}E[N_P]\ /\ E[Y] \qquad (6)$$

It is easy to realise that

$$Y = \sum_{i=1}^{K_w-1}\Big(C_I + C_B(|failure)\Big) + C_I + C_B(|success) \qquad (7)$$

where K_w is the number of transmission attempts needed until success. For a system without a capture effect, any overlapping transmission will be destructive and $p_{net} = P\ (K_{Tx} = 1)$. However, with the spread-spectrum radio ($B_k > 0$) the first out of k concurrent transmissions may succeed, but if and only if the first transmission is not addressed to one of the other transmitting nodes. The first condition is given by B_k (4) while the second, the probability that the first transmission out of k is destined for a receiving node, is given by $(n - k)$ / $(n - 1)$. Hence,

$$p_{net} = \sum_{k=1}^{n}\frac{n-k}{n-1}\cdot B_k \cdot P\big(K_{Tx} = k\big) \qquad (8)$$

The transmission attempts are independent and then K_w is geometrically distributed with the mean $1/p_{net}$. From (7) we find the first moment of Y as

$$\begin{aligned} E[Y] &= (1/p_{net} - 1)(E[C_I] + E[C_B|failure]) + E[C_I] \\ &\quad + E[C_B|success] = E[C_I]\ /\ p_{net} \\ &\quad + \{(1 - p_{net})E[C_B|failure] \\ &\quad + p_{net}E[C_B|success]\}\ /\ p_{net} \\ &= (E[C_I] + E[C_B])\ /\ p_{net} \end{aligned} \qquad (9)$$

and then (6) becomes

$$\lambda_{max} = p_{net}t_{bz}E[N_P]/(E[C_I] + E[C_B]) \qquad (10)$$

The pdf for K_{Tx} is the subject of section 3.1 while $E[C_I]$ and $E[C_B]$ are treated in section 3.2. When all the unknowns have been resolved, λ_{max} will be validated by simulation.

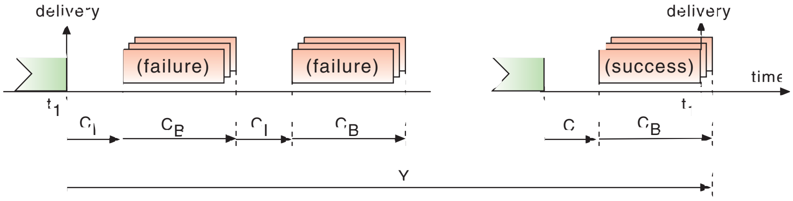


Figure 4 The delivery cycle Y as seen by an outside observer

3.1 The number of simultaneous transmissions

We consider the system at a time t_1 when a transmission has just ended, see Figure 5. A MAC entity that has a packet ready for transmission will draw its random access delay and starts to transmit at its expiration time if no other transmission is detected in between. A MAC entity that is idle at t_1 may get a packet some time later and then takes the same sequence of actions as the other MAC entities. The MAC protocol parameter t_u (5) has a great influence on the network performance. As t_u decreases the probability that two or more nodes transmit simultaneously increases, while the time passed until the first transmission occurs decreases. As t_u increases the opposite happens. At a given user traffic and a set of system parameters, a t_u value exists that gives an optimum balance between the channel idle time and the channel capacity wasted due to collisions.

We shall attempt to find the probability of having exactly k simultaneous transmissions during the contention period, but first we need some definitions:

$n \overset{\Delta}{=}$ the number of nodes in the network

$T_\Lambda \overset{\Delta}{=}$ the MAC SDU inter arrival time

$c_i \overset{\Delta}{=}$ the probability that a MAC entity has a MAC PDU to send at a random point in time given that it is not transmitting

$S \overset{\Delta}{=} T_\Lambda + D$; the MAC entity scheduling interval

T_Λ is assumed to be negative exponentially distributed with the expectation $1/\Lambda_i$. The total packet arrival intensity on the radio channel is $\Lambda \overset{\Delta}{=} n \cdot \Lambda_i$.

Among others, c_i depends on Λ and we shall later derive a set of equations that can be used to calculate c_i. If a MAC entity has a packet to send (is busy) at t_1, then the scheduling interval has the cdf $P(S \leq t \mid \mathit{busy\ at}\ t_1) = F_D(t)$. If a MAC entity is idle at t_1 then the scheduling interval is given by the convolution

$$P(S \leq t \mid \mathit{idle\ at}\ t_1)$$
$$= \int_0^\infty \left(1 - e^{-\Lambda_i(t-x)}\right) f_D(x)dx = F_D(t) - e^{-\Lambda_i t} \int_{t_l}^t e^{\Lambda_i x} f_D(x)dx$$

where

$$e^{-\Lambda_i t} \int_{t_l}^t e^{\Lambda_i x} f_D(x)dx$$

$$= \begin{cases} \frac{1}{\Lambda_i(t_u - t_l)} \left[1 - e^{-\Lambda_i(t-t_l)}\right] & \text{for } t \in [t_l, t_u] \\ \frac{1}{\Lambda_i} \frac{e^{\Lambda_i t_u} - e^{\Lambda_i t_l}}{(t_u - t_l)} e^{-\Lambda_i t} & \text{for } t \geq t_u \end{cases} \qquad (11)$$

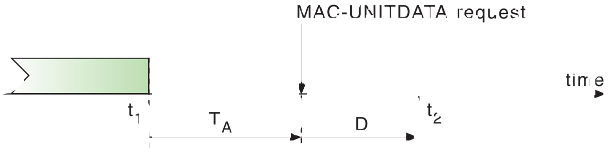


Figure 5 A MAC entity idle at t_1 becomes busy at $t_1 + t_\Lambda$ and draws a random access delay D. If no transmission is detected within the period D, the MAC entity enters the transmission state at t_2

Total probability gives

$$P(S \leq t) = F_D(t) - (1 - c_i)e^{-\Lambda_i t} \int_{t_l}^t e^{\Lambda_i x} f_D(x)dx \qquad (12)$$

Hence, for $\Lambda_i > 0$

$$F_S(t) = \begin{cases} 0 & \text{for } t \leq t_l \\ \frac{t-t_l}{t_0} - \frac{1-c_i}{\Lambda_i t_0}\left[1 - e^{-\Lambda_i(t-t_l)}\right] & \text{for } t \in [t_l, t_u] \\ 1 - \frac{1-c_i}{\Lambda_i} \cdot \frac{e^{\Lambda_i t_u} - e^{\Lambda_i t_l}}{t_0} e^{-\Lambda_i t} & \text{for } t \geq t_u \end{cases} \qquad (13)$$

where $t_0 \overset{\Delta}{=} t_u - t_l$. This cdf for S turns out to give some non-trivial expressions. We therefore use a linear approximation in the range $[t_l, t_u]$ and get

$$F_S(t) = \begin{cases} 0 & \text{for } t \leq t_l \\ \frac{C_0}{t_0}(t - t_l) & \text{for } t \in [t_l, t_u] \\ 1 - (1 - C_0)e^{-\Lambda_i(t-t_u)} & \text{for } t \geq t_u \end{cases} \qquad (14)$$

where $C_0 \overset{\Delta}{=} P(S \leq t_u)$ given by (13).

Figure 6 plots $F_S(t)$ for some values of c_i and Λ_i. The deviation between the exact result and the approximation increases with increasing Λ_i and decreasing c_i. When $c_i = 1$ they are identical.

Let $S_1 \leq S_2 \leq ... \leq S_n$ be the order statistics obtained by ordering the scheduling intervals. The node that gets the smallest scheduling interval starts to transmit after the delay S_1, see Figure 7.

If $S_1 = t$ and $S_k \in [t, t + t_v]$ then at least k simultaneous transmissions have occurred. The probability that a particular node

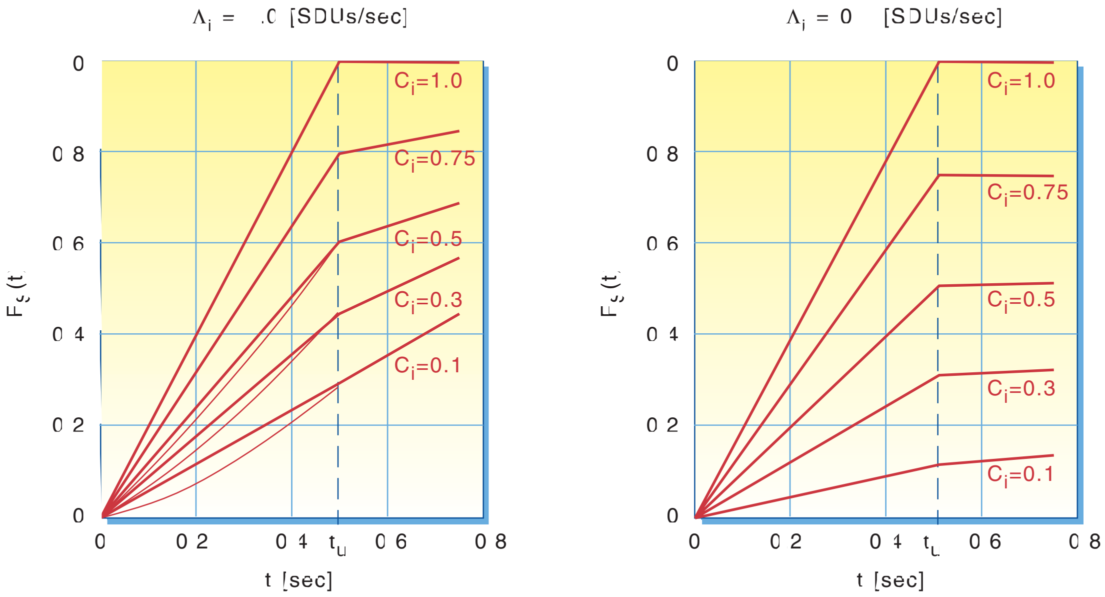


Figure 6 Numerical values for $F_S(t)$ for different c_i and Λ_i. The approximation as thin lines and the exact result as thick lines. t_l = 0 and t_u = 0.5 sec

transmits within $[t, t + t_v]$ given that it did not transmit before t, is given by

$$a(t) = \frac{F_S(t+t_v) - F_S(t)}{1 - F_S(t)} \tag{15}$$

Any of the $(n - 1)$ remaining nodes may start an additional transmission within the vulnerable period of the first, and we assume that they all behave independently. Then K_{Tx} is binomially distributed and

$$p_{k,n}(t) \overset{\Delta}{=} P(K_{Tx} = k \mid \textit{the first of n nodes starts to transmit at t})$$

$$= \binom{n-1}{k-1}[a(t)]^{k-1}[1-a(t)]^{n-k} \tag{16}$$

and the pgf (probability generating function) of $p_{k,n}(t)$ is

$$p_n(z,t) = \sum_{k=1}^{n} p_{k,n}(t)z^k = z[z \cdot a(t) + 1 - a(t)]^{n-1} \tag{17}$$

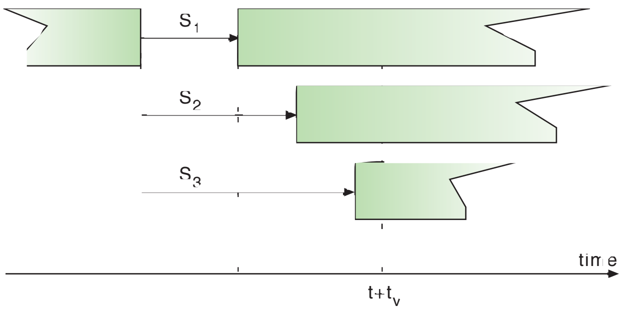


Figure 7 The winning node transmits at t

The probability that one of the n nodes starts a transmission at t is given by

$$\frac{d}{dt}\left[1-\left(1-F_S(t)\right)^n\right] = n\left(1-F_S(t)\right)^{n-1} f_S(t).$$

Substituting (15) into (17) and unconditioning give

$$p_n(z) =$$

$$n \cdot z \int_0^\infty \left\{ z\left[F_S\left(t+t_v\right) - F_S(t)\right] + 1 - F_S\left(t+t_v\right)\right\}^{n-1} f_S(t)dt \tag{18}$$

The approximation (14) for the scheduling distribution gives

$$p_n(z) = n \cdot z \int_{t_l}^{(t_u - t_v)} \left\{ zC_0\frac{t_v}{t_0} + 1 - \frac{C_0}{t_0}\left(t + t_v - t_l\right)\right\}^{n-1} \cdot \frac{C_0}{t_0} dt + z \cdot$$

$$n\Lambda_i\left(1-C_0\right)^n e^{n\Lambda_i t_u} \cdot \left[z\left(1-e^{-\Lambda_i t_v}\right) + e^{-\Lambda_i t_v}\right]^{n-1} \int_{(t_u - t_v)}^{\infty} e^{-n\Lambda_i t} dt \tag{19}$$

where we in the second integral have introduced yet another approximation when $t_v << t_0$ by using the $1-(1-C_0)e^{-\Lambda_i(t-t_u)}$ component of the $F_S(t)$ in the range $(t_u - t_v) \le t \le t_u$. Integration gives

$$p_n(z) = z\left[zC_0\frac{t_v}{t_0} + 1 - C_0\frac{t_v}{t_0}\right]^n - z\left[zC_0\frac{t_v}{t_0} + 1 - C_0\right]^n +$$

$$z\left(1-C_0\right)^n \cdot \left\{ z\left[1-e^{-\Lambda_i t_v}\right] + e^{-\Lambda_i t_v}\right\}^{n-1} e^{n\Lambda_i t_v} \tag{20}$$

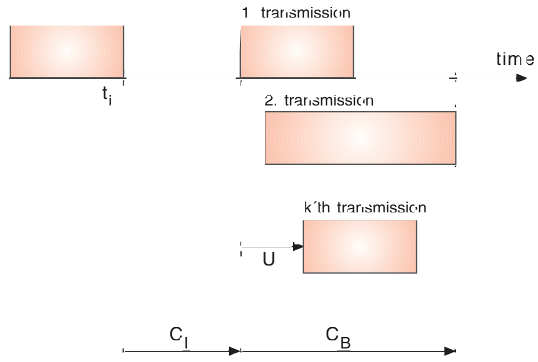


Figure 8 A transmission ends at t_l and the time delay C_I passes before the first transmission starts. The channel will stay busy for a time C_B

Series expansion of $p_n(z)$ gives

$$P(K_{Tx}=k)=\binom{n}{k-1}\left(C_0\frac{t_v}{t_0}\right)^{k-1}\left(\left(1-C_0\frac{t_v}{t_0}\right)^{n-k+1}-(1-C_0)^{n-k+1}\right)+\binom{n-1}{k-1}(1-C_0)^n\left(1-e^{-\Lambda_i t_v}\right)^{k-1}e^{k\Lambda_i t_v} \quad (21)$$

for $1 \leq k \leq n$, $t_l < t_u$, $t_v < t_0$ and $n > 1$.

(21) gives a relationship between K_{Tx} and the user traffic through C_0. C_0 is a function of c_i and Λ_i. At low load c_i, $\Lambda_i \approx 0$ and hence $C_0 \approx 0$, and we find that $P(K_{Tx} = 1) = 1$ and $P(K_{Tx} = k) = 0$ for $k \geq 2$. Only one node will transmit at a time. The same result is achieved by setting $t_v = 0$ in (21); $t_v = 0$ gives a collision-free network regardless of the offered traffic. Under maximum load, i.e. $c_i = 1$, (21) becomes

$$P\left(K_{Tx}=k\right)=\binom{n}{k-1}\left(\frac{t_v}{t_0}\right)^{k-1}\left(1-\frac{t_v}{t_0}\right)^{n-k+1} \quad (22)$$

and we note that

$$\sum_{k=1}^{n}P\left(K_{Tx}=k\right)=p_n\left(z|C_0=1\right)=1-\left(t_v/t_0\right)^n$$

This sum should actually be one independent of t_v at $c_i = 1$. The deviation is due to the approximation used in the second integral of (19).

3.2 The channel idle and busy period

The objective of this chapter is to find expressions for the channel idle period and the channel busy period. A transmission ends at t_1 in Figure 8 and the time delay to the first transmission is C_I, the channel idle period. If none of the network nodes have a ready packet when a transmission finishes, a time delay will pass until the first node gets one. Then the node will draw a random access delay.

A channel busy period C_B starts when the *first* radio is turned into transmission mode and ends when the *last* radio, of those transmitting during this channel access, has returned to the idle state.

The cdf for the C_I is given by

$$P(C_I \leq t) = P(\mathrm{Min}[S_1, ..., S_n] \leq t) = 1 - [1 - F_S(t)]^n$$

giving

$$E\left[C_I\right]=\int_0^\infty\left[1-P\left(C_I\leq t\right)\right]dt=\int_0^\infty\left[1-F_S(t)\right]^n dt$$

The scheduling approximation (14) gives

$$E\left[C_I\right]=\int_0^{t_l}dt+\int_{t_l}^{t_u}\left[1-\frac{C_0}{t_0}\left(t-t_l\right)\right]^n dt+\left(1-C_0\right)^n e^{\Lambda_i n t_u}\int_{t_u}^{\infty}e^{-\Lambda_i n t}dt$$
$$=t_l+\frac{t_0}{(n+1)C_0}+\left(1-C_0\right)^n\left(\frac{1}{n\Lambda_i}-1+C_0\right) \quad (23)$$

Now we shall find the cdf for the busy period in case of stochastic packet length given by (1) when $N_p = DiscreteUniform$ [1, ..., a]. We define:

$V \overset{\Delta}{=}$ the transmission end-point for an arbitrary node that begins its transmission after the first (the transmissions 2 ... k in Figure 8).

$U \overset{\Delta}{=}$ the delayed start for an arbitrary node that transmits within the vulnerable period of the first.

We find the cdf $F_V(t)$ in the usual way by convoluting the delayed start distribution and the packet length distribution:

$$F_V(t)=\int_0^t F_{T_p}(t-\tau)\cdot f_U(\tau)d\tau \quad (24)$$

Competing transmissions may arrive in the vulnerable period of the first. Given that a transmission hits the vulnerable period, the hit point will be uniformly distributed in that interval. If $C_0 < 1$ and $C_I = t$, the interval is $[t, t + t_v]$. However, if $C_0 = 1$ the interval will be $[t, t_u - 1]$ for $t_u - t_v \leq t \leq t_u$. When $t_v << t_0$ we can use an approximation neglecting this fact. Therefore

$$F_U(t) = t / t_v \qquad \text{for } 0 \leq t \leq t_v \quad (25)$$

Figure 9 shows the convolution (24) when $t_v \leq t_{bz}$. The case $t_v > t_{bz}$ is of little practical interest because it means that concurrent transmissions within a contention period may be separated by a time gap (of maximum size $t_v - t_{bz}$) with no “carrier” on the radio channel. By inspecting Figure 9 we see that

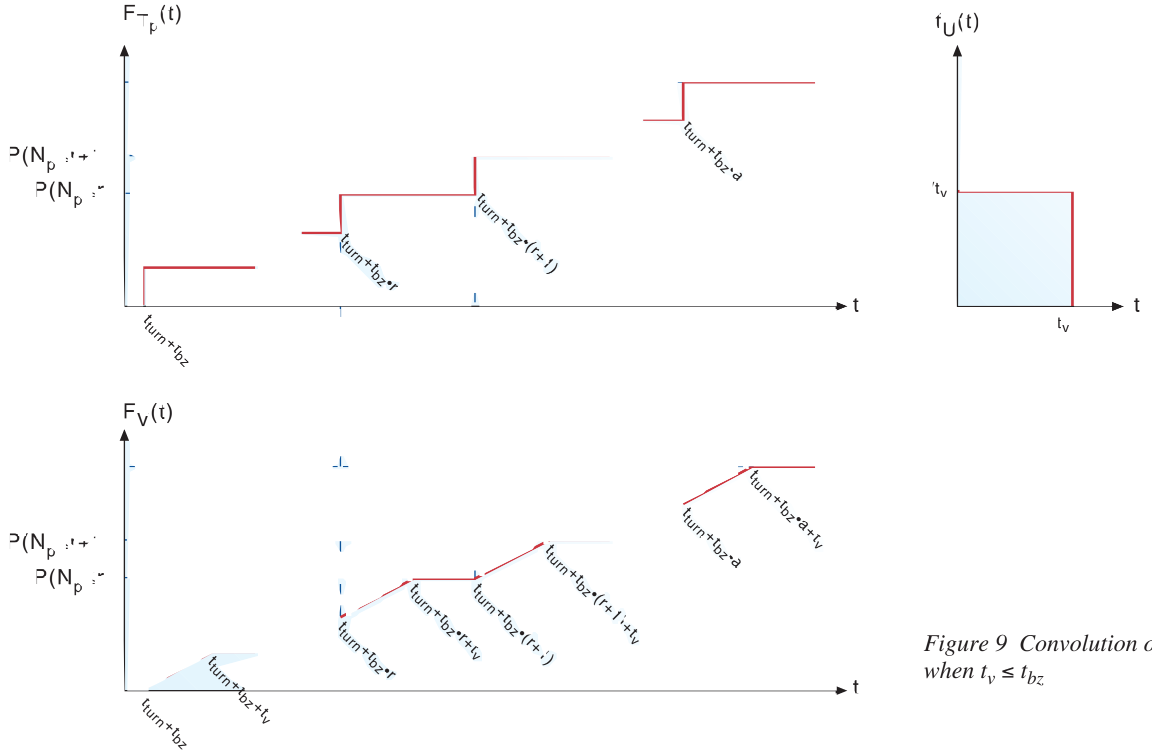


Figure 9 Convolution of T_p and U when $t_v \le t_{bz}$

$$F_v(t) = \begin{cases} P(N_p < r) + \frac{t - t_{turn} - r\cdot t_{bz}}{t_v} P(N_p = r) \\ \text{for } t_{turn} + r\cdot t_{bz} \le t \le t_{turn} + r\cdot t_{bz} + t_v \\ P(N_p \le r) \\ \text{for } t_{turn} + r\cdot t_{bz} + t_v \le t \le t_{turn} + (r+1)\cdot t_{bz} \end{cases}$$

Then

$$P(C_B \le \tau | C_I = t, K_{Tx} = k)$$
$$= P\left(Max\left[Tp, \overbrace{V, ..., V}^{k-1} \right] \le \tau \right) = F_{Tp}(\tau)\left[F_V(\tau)\right]^{k-1} \tag{26}$$

The cdf for the T_p is simply $P(N_p \le r)$ for $t_{turn} + r\cdot t_{bz} < t \le t_{turn} + (r+1)\cdot t_{bz}$. Unconditioning on k gives

$$P(C_B \le \tau | C_I = t) =$$
$$F_{T_p}(\tau)\sum_{k=1}^{n}\left[F_V(\tau)\right]^{k-1} P(K_{Tx} = k | C_I = t)$$

By comparing with equation (17) we see that

$$P(C_B \le \tau | C_I = t) = F_{T_p}(\tau)\cdot p_n(F_V(\tau), t)/F_V(\tau)$$

and unconditioning on t by (18) gives

$$P(C_B \le \tau) = F_{T_p}(\tau)\cdot p_n(F_V(\tau))/F_V(\tau)$$

By (20) we have

$$P(C_B \le t) = F_{T_p}(t)\cdot$$
$$\left\{ \begin{array}{l} \left[F_V(t)C_0\frac{t_v}{t_0} + 1 - C_0\frac{t_v}{t_0}\right]^n - \left[F_V(t)C_0\frac{t_v}{t_0} + 1 - C_0\right]^n \\ +(1-C_0)^n\cdot\left\{F_V(t)\left[1 - e^{-\Lambda_i t_v}\right] + e^{-\Lambda_i t_v}\right\}^{n-1} e^{n\Lambda_i t_v} \end{array} \right\} \tag{27}$$

From this we may find the moments of the C_B in the usual way, but the expressions get lengthy so instead we present the conditional moments (in the forthcoming delay analysis the second moment is needed).

$$E[C_B | K_{Tx} = k] = \int_0^\infty [1 - P(C_B \le t | K_{Tx} = k)]$$

$$= \sum_{r=1}^{a}\int_0^\infty \{1 - F_{T_p}(t)[F_V(t)]^{k-1}\}dt = t_{turn} + t_{bz}\sum_{r=0}^{a}\{1 - [P(N_p \le r)]^k\}$$

$$+t_v\sum_{r=1}^{a}\left\{[P(N_p \le r)]^k - \frac{P(N_p \le r)}{kP(N_p = r)}\cdot\left([P(N_p \le r)]^k - [P(N_p < r)]^k\right)\right\} \tag{28}$$

$$E[C_B^2|K_{Tx}=k] = 2\int_0^\infty t[1-P(C_B \le t|K_{Tx}=k)]dt$$

$$= (t_{turn}+t_{bz})^2 + t_{bz}\sum_{r=1}^{a}(2t_{turn}+2r\cdot t_{bz}+t_{bz})\cdot\left[1-[P(N_p\le r)]^k\right]$$

$$+t_v\sum_{r=1}^{a}(2r\cdot t_{bz}+t_v)[P(N_p\le r)]^k + 2a\cdot t_{turn}\cdot t_v - 2\frac{t_v}{k}\sum_{r=1}^{a}\frac{P(N_p\le r)}{(N_p=r)}\cdot$$

$$\left\{(t_{turn}+r\cdot t_{bz})\cdot\left([P(N_p\le r)]^k-[P(N_p<r)]^k\right)+t_v\,[P(N_p\le r)]^k\right\}$$

$$+2\frac{t_v^2}{k(k+1)}\sum_{r=1}^{a}\frac{P(N_p\le r)}{P(N_p=r)^2}\cdot\left\{[P(N_p\le r)]^{k+1}-[P(N_p<r)]^{k+1}\right\} \tag{29}$$

In our case $P(N_p \le r) = r/a$, $P(N_p < r) = (r-1)/a$ and $P(N_p = r) = 1/a$, and (28) and (29) can be written as

$$E[C_B|K_{Tx}=k] = t_{turn}+t_{bz}\sum_{r=0}^{a}\left(1-\left(\frac{r}{a}\right)^k\right)$$

$$+t_v\sum_{r=1}^{a}\left(\frac{r}{a}\right)^k\frac{k-r+r(1-1/r)^k}{k} \tag{30}$$

$$E[C_B^2|K_{Tx}=k] = (t_{turn}+t_{bz})^2 + t_{bz}\sum_{r=1}^{a}(2t_{turn}+2r\cdot t_{bz}+t_{bz})\cdot$$

$$[1-(r/a)^k]+2a\cdot t_{turn}\cdot t_v + 2t_v\sum_{r=1}^{a}r\left(\frac{r}{a}\right)^k\cdot$$

$$\left\{t_{bz}-\frac{t_{turn}+r\cdot t_{bz}}{k}(1-(r-1)^k/r^k)\right\}+t_v^2\sum_{r=1}^{a}\left(\frac{r}{a}\right)^k\cdot$$

$$\left\{1-2\frac{r}{k}+2\frac{r}{k(k+1)}\left(r-(r-1)^{k+1}/r^k\right)\right\} \tag{31}$$

and by unconditioning we find the moments of the C_B as

$$E[C_B^m] = \sum_{k=1}^{n}E[C_B^m|K_{Tx}=k]P(K_{Tx}=k) \quad \text{for } m = 1,2 \tag{32}$$

where $P(K_{Tx} = k)$ is given by (21). (30) and (31) give

$$E[C_B|K_{Tx}=k,a=1] = t_{turn}+t_{bz}+\frac{k-1}{k}t_v$$

$$E[C_B^2|K_{Tx}=k,a=1] = (t_{turn}+t_{bz})^2$$

$$+2t_v(t_{turn}+t_{bz})\frac{k-1}{k}+t_v^2\cdot\frac{k-1}{k+1}$$

The results for other fixed packet lengths follow directly by substituting $t_{turn} + t_{bz}$ by the true length, leading to the much simpler expressions

$$E\left[C_B\middle|\text{fixed}T_p\right] = t_v\sum_{k=1}^{n}\frac{k-1}{k}P\left(K_{Tx}=k\right)+T_p$$

$$E\left[C_B^2\middle|\text{fixed}T_p\right] = t_v^2\sum_{k=1}^{n}\frac{k-1}{k+1}P\left(K_{Tx}=k\right)$$

$$+2T_pt_v\sum_{k=1}^{n}\frac{k-1}{k}P\left(K_{Tx}=k\right)+T_p^2$$

3.3 Case study

This section validates the throughput model (6) by simulating a few cases using the layer parameter setting in Table 2.

Table 2 Layer parameter values

Block Size t_{bz}	48 msec
t_{turn}	3.6 msec
t_{sync}	6.4 msec
capture model $B_k = q^{k-1}$	$q = 0.82$
t_l	0

Figure 10 plots the throughput by (10) together with simulated results and shows excellent agreement both for low and high collision rates. The highest and lowest collision rate tested occur at ($n = 25$, $t_u = 0.1$) and $n = 9$, $t_u = 1.0$, giving $E[K_{Tx}] = 3.5$ and $E[K_{Tx}] = 1.1$, respectively. Small t_u-values lead to a short average channel idle period but throughput degradation occurs as a result of packet loss; on the first hand determined by $SNR_{capture}$ and then on the probability of addressing another transmitting node. By increasing t_u, fewer collisions occur but the average channel idle period is also increased. For a set of system parameters and packet lengths, a t_u-value exists that maximises the throughput.

4 The MAC and LLC service times

We shall estimate the network transit delay by a queuing model for each node. The first step is to find expressions for the moments of the MAC service time represented by the stochastic variable X in Figure 11. X is the time a MAC entity uses to grab the channel and transmit one MAC SDU. An LLC SDU that arrives to an empty queue will not experience a queue delay Q, but may instead be delayed a time H – the residual channel busy period at a random point in time. However, an LLC SDU that arrives to a non-empty queue gets a queuing delay and not the delay H. This special behaviour is taken into account by modelling each network node as an M/G/1-queue with set-up times as reported in [7]. The mean time spent in the LLC queue follows directly from [8] (2.44a):

$$E[Q] = \frac{2E[H]+\Lambda_iE\left[H^2\right]}{2\left(1+\Lambda_iE[H]\right)}+\frac{\Lambda_iE\left[X^2\right]}{2\left(1-\Lambda_iE[X]\right)} \tag{33}$$

When H has passed, the MAC entity and all its competitors draw a random access delay. The MAC service time X is composed of the following stochastic variables:

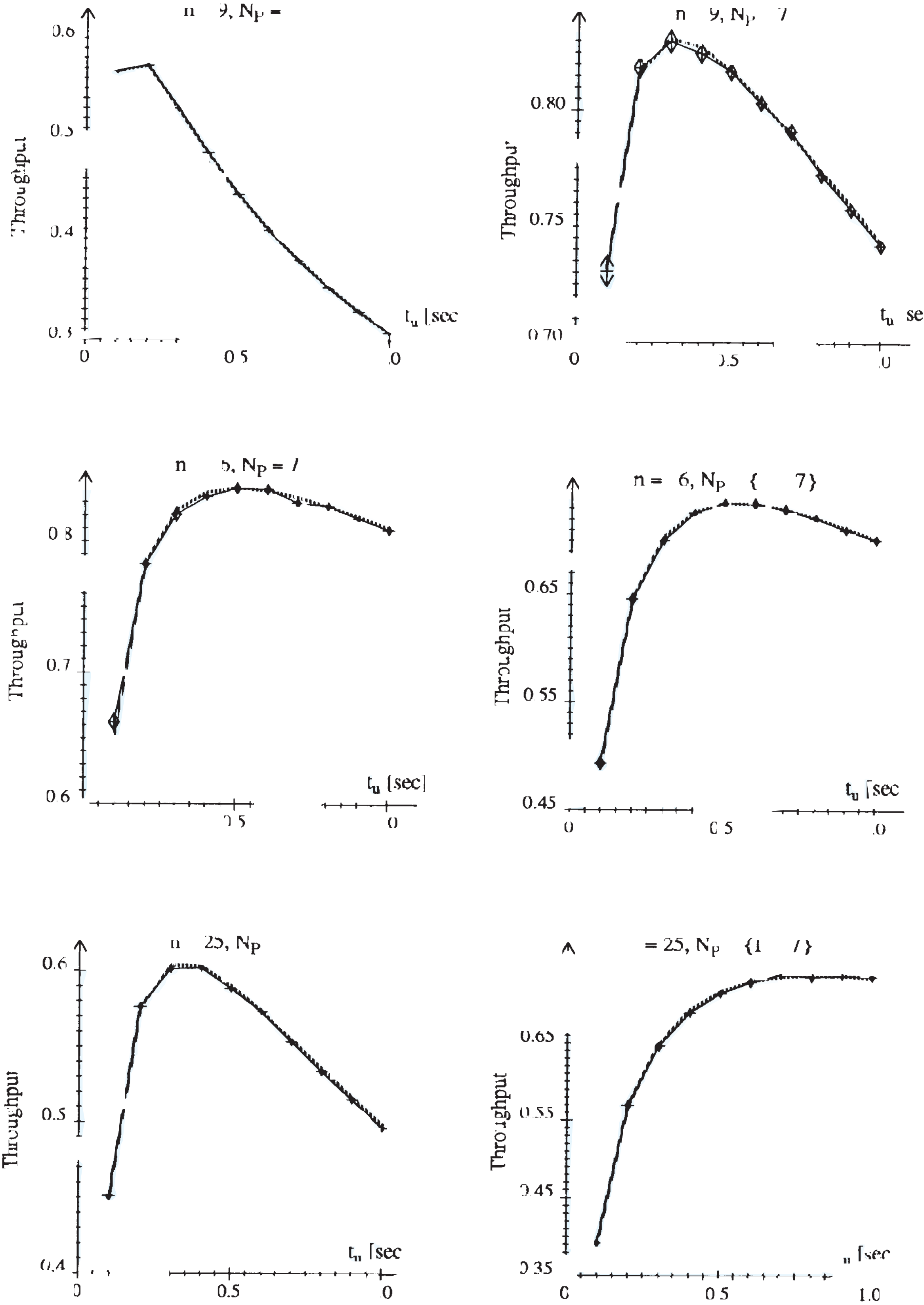


Figure 10 Simulated and theoretical normalised throughput as function of t_u. Simulated results as solid lines with 90 % confidence intervals

$K_w \triangleq$ the number of contention periods passed before the node transmits

$L \triangleq$ the channel idle period seen by a busy node given that it does not transmit in the following busy period (the node loses)

$W \triangleq$ the channel idle period seen by a busy node given that it does transmit in the following busy period (the node wins)

$B \triangleq$ the channel busy period seen by a busy node given that it does not transmit.

We assume that the stochastic variables are independent and that their individual distributions remain identical from contention period to contention period. We consider one tagged busy node and by inspecting Figure 11 we have

$$X = \sum_{j=1}^{K_w-1} (L+B) + W + T_p \tag{34}$$

Hence

$$E[X] = \{E[L] + E[B]\}E[K_w - 1] + E[W] + E[T_p] \tag{35}$$

$$\begin{aligned} E[X^2] = & E[K_w - 1]\mathrm{Var}(L + B) + \{E[L] + E[B]\}^2\mathrm{Var}(K_w) \\ & + \mathrm{Var}(W) + \mathrm{Var}(T_p) + \{E[X]\}^2 \end{aligned} \tag{36}$$

We point out that L and B are dependent. The probability that one or more of n nodes start to transmit at t is given by $\Phi_n(t) = 1 - [1 - F_S(t)]^n$ and the probability σ that the tagged busy node transmits in a contention period is

$$\sigma = \int_0^\infty f_D(t)[1 - \Phi_{n-1}(t - t_v)]dt \tag{37}$$

where the bracketed sum is the probability that none of its $n-1$ neighbours transmit before $t - t_v$. Any node will detect a busy radio channel t_v after the first transmission has started. Inserting the approximation (14) into (37) gives

$$\begin{aligned} \sigma &= \int_{t_l}^{(t_v+t_l)} f_D(t)dt + \int_{(t_v+t_l)}^{t_u} f_D(t)[1 - F_S(t - t_v)]^{n-1}dt \\ &= \frac{t_v}{t_0} + \frac{1}{nC_0}\left[1 - \left(1 - C_0 + C_0\frac{t_v}{t_0}\right)^n\right] \end{aligned} \tag{38}$$

By our independence assumption K_w is geometrically distributed with the mean $1/\sigma$. The pdf for the conditional stochastic variable W follows directly by the arguments leading to (37)

$$\begin{aligned} f_W(t) &= \frac{f_D(t)[1 - \Phi_{n-1}(t - t_v)]}{\sigma} \\ &= \frac{f_D(t)[1 - F_S(t - t_v)]^{n-1}}{\sigma} \end{aligned} \tag{39}$$

The tagged busy node does not transmit at t if some of its neighbour transmits at $t - t_v$ and the pdf for the conditional stochastic variable L is

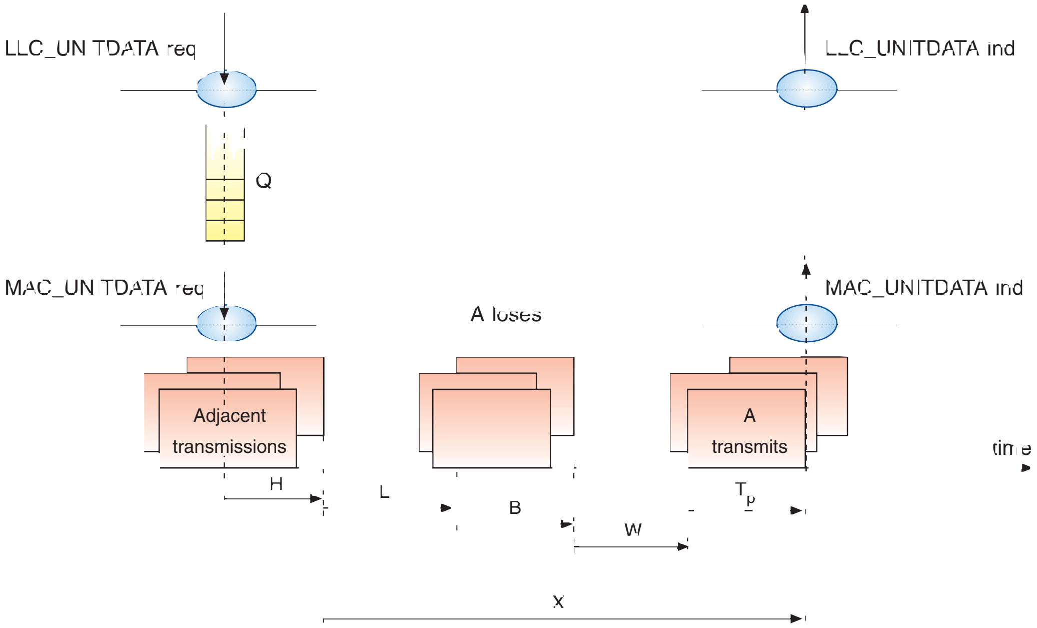


Figure 11 One possible sequence of events until a particular node A has grabbed the channel and sent a packet. If the transmission was successful, the destination LLC entity receives a MAC-UNITDATA indication

$$f_L(t) = \frac{1-F_D(t)}{1-\sigma}\frac{d}{dt}\Phi_{n-1}(t-t_v)$$

$$= \frac{n-1}{1-\sigma}[1-F_S(t-t_v)]^{n-2}\cdot[1-F_D(t)]f_S(t-t_v) \tag{40}$$

We need the first and second moments of W and L given by

$$E[W^m] = \frac{1}{\sigma}\int_{t_l}^{(t_l+t_v)} t^m \cdot f_D(t)dt$$

$$+\frac{1}{\sigma}\int_{(t_l+t_v)}^{t_u} t^m \cdot f_D(t)[1-F_S(t-t_v)]^{n-1}dt \tag{41}$$

$$E[L^m] = \frac{n-1}{1-\sigma}\int_{(t_l+t_v)}^{t_u} t^m \cdot [1-F_S(t-t_v)]^{n-2}\cdot$$

$$[1-F_D(t)]f_S(t-t_v)dt \tag{42}$$

These integrals are easily solved but the results contain many terms and are therefore not included. To find $E[(L+B)^2]$ under dependence, we condition on that the first transmission starts at t and under this condition we seek the cdf $F_B(t)$. We observe that the possible overlaps are generated by $(n-1)$ nodes and in accordance with equation (17), the pgf for the number of overlaps is $p_{n-1}(z, t) = z[z \cdot a(t) + 1 - a(t)]^{n-2}$. When the vulnerable period is much smaller than the packet length we have $P(B \geq t$ | *the first of n nodes starts to transmit at t*) =

$$\sum_{k=1}^{n-1}[F_{T_p}(\tau)]^k = p_{n-1}(F_{T_p}(\tau), t)$$

$$= F_{T_p}(\tau)[F_{T_p}(\tau)\cdot a(t) + 1 - a(t)]^{n-2}$$

To find $E[(L+B)^2]$ from here is possible but lengthy.

For an increasing number of nodes the resulting scheduling process will tend to a Poisson process. This means that the probability distribution for the number of simultaneous transmissions in the interval $[t, t+t_v]$, the first transmission starts at t, will be independent of t. Since B only depends on t through this distribution, L and B will in the limit be independent. Under this assumption we can use $E[C_B^2 \mid K_{Tx} = k]$ as given by (31) for $(n-1)$ nodes. A busy node that does not transmit will experience a busy period t_v shorter than the outside observer because the "carrier" sense is delayed t_v after the start of the first transmission. Hence

$$B(|K_{Tx}^{\bigtriangledown} = k) = C_B(|K_{Tx} = k) - t_v$$

where $K_{Tx}^{\bigtriangledown}$ is the number of transmissions seen by the busy node. By finding the first and second moments of

$B(|K_{Tx}^{\bigtriangledown} = k)$ and then unconditioning we have

$$E[B] = \sum_{k=1}^{n-1} E[C_B|K_{Tx}=k]P\left(K_{Tx}^{\bigtriangledown}=k\right) - t_v \tag{43}$$

$$E[B^2] = \sum_{k=1}^{n-1} E[C_B^2|K_{Tx}=k]P\left(K_{Tx}^{\bigtriangledown}=k\right)$$

$$-2t_vE[B] - t_v^2 \tag{44}$$

where $P(K_{Tx}^{\bigtriangledown}=k)$ is given by (21) when n is substituted by $n-1$.

The moments of the residual channel busy period at a random point in time is given by [9] (5.14):

$$E[H^m|\text{the channel is busy}] = \frac{E[B^{m+1}]}{(m+1)E[B]} \tag{45}$$

The probability of an ongoing transmission, seen from a particular node, is $\Lambda_i(n-1)E[B]$. Removing the condition

$$E[H^m] = \Lambda_i(n-1)E[B]\frac{E[B^{m+1}]}{(m+1)E[B]} = \Lambda_i(n-1)\frac{E[B^{m+1}]}{(m+1)} \tag{46}$$

To evaluate $E[H^2]$ under stochastic packet length the third moment of the B is needed, while we have $E[B]$ and $E[B^2]$ only. We neglect the variation of H by using the approximation $E[H^2] = (E[H])^2$ in this case.

A node is said to be busy when it is waiting, scheduling or transmitting, i.e. it has a packet to send. Viewed as a queuing system [8] (2.46b) gives

$$P\ (node\ \ i\ \ busy) = \frac{E[X]+E[H]}{1+\Lambda_i E[H]}\Lambda_i \tag{47}$$

for the M/G/1-queue with set-up times. Under stationary conditions then

$$c_i \triangleq P(\text{node } i \text{ busy} \mid \text{it does not transmit})$$

$$= \frac{P(\text{node } i \text{ busy}) - \Lambda_i E[T_p]}{1-\Lambda_i E[T_p]} \tag{48}$$

By (48) we have set of non-linear equations that may be solved numerically to find c_i. The approach neglects dependencies among the various nodes except from those through the c_i's.

4.1 Case study

Now we compare theoretical and simulated results for $n \in \{9, 16, 25\}$ and use the same LLC SDU sizes as done under

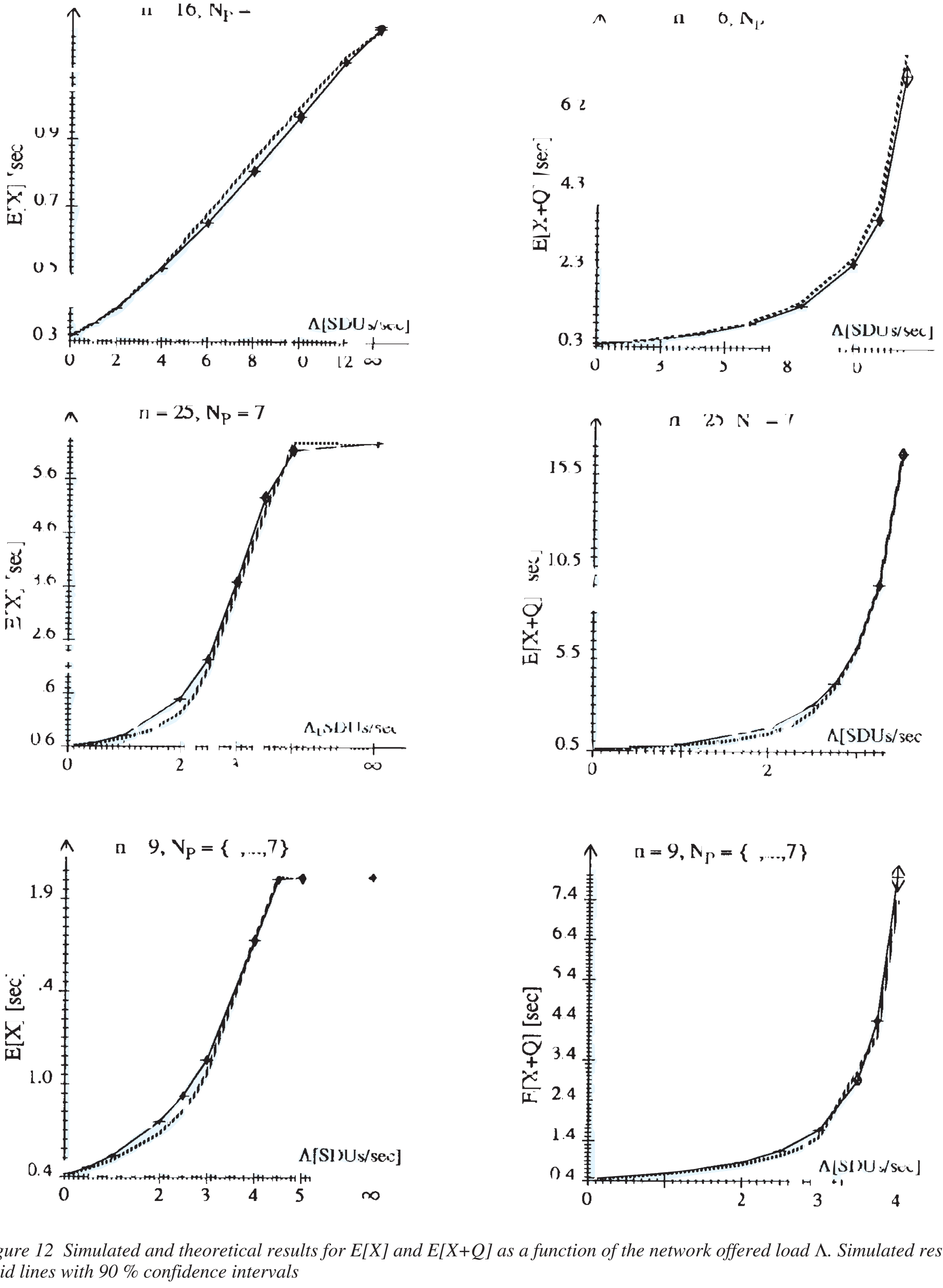


Figure 12 Simulated and theoretical results for E[X] and E[X+Q] as a function of the network offered load Λ. Simulated results as solid lines with 90 % confidence intervals

the previous throughput simulations. The throughput simulations showed that $t_u = 0.5$ sec is a reasonable value to use. Figure 12 shows six of the eighteen performance curves (the other twelve show the same accuracy).[1] c_i increases with increasing offered load and when c_i reaches one, the MAC service time has reached its limit. The LLC service time has, of course, no limit.

5 Conclusion

In this paper, we have analysed the performance of a single-channel spread-spectrum radio network using a random access protocol. In spread-spectrum radio networks even packets that suffer from strong interfering sources, e.g. caused by the collisions that occur when using random access protocol, have good chance of being correctly received. This is determined by the radio demodulation probabilities under different signal-to-noise ratios. An exact modelling of a real radio leads to intractable network performance models, and we presented a method that gives a fair accuracy.

Parts of the theoretical modelling assumes that t_v / t_0 is small. We showed that this will be true for real networks with reasonable t_u-values, that is, values not giving significant throughput degradation. In the validation of the throughput model at maximum load, we used tu-values such that $0.01 \leq t_v / t_0 \leq 0.1$ and found excellent conformity between theoretical and simulated results.

The network transit delay is the sum of two components; the MAC service time and the queuing delay within the LLC layer. The queuing delay was found under some independent assumptions considering each node separately. By comparing with simulation results we showed that this approach gives good accuracy for the network transit delay under the assumptions given.

[1] *One problem encountered when estimating E[Q] by taking samples from the waiting time in queue is the correlation between the samples. The technique used to combat the correlation problem is the batch means method [2]. The simulation is first terminated when the lag 1 autocorrelation becomes less than 0.2 (at a 90 % confidence level).*

References

1 Pickholtz, R L. Theory of spread-spectrum communications : a tutorial. *IEEE transactions on communications,* Com-30(5), 1982.

2 Pawlikowski, K. Steady-state simulation of queueing processes : a survey of problems and solutions. *ACM computing surveys,* 22(2), 1990.

3 Kleinrock, L, Lam, S. Packet switching in a multiaccess broadcast channel : performance evaluation. *IEEE transactions on communications,* Com-23, 1975.

4 Tobagi, F A. Analysis of a two-hop centralized packet radio network, part II : carrier sense multiple access. *IEEE transactions on communications,* Com-28, 1980.

5 Boorstyn, et al. Throughput analysis in multiple CSMA packet radio networks. *IEEE transactions on communications,* Com-35, 1987.

6 Yu, H-C, Hamilton, R L. A buffered two-node packet radio network with product form solution. *IEEE transactions on communications,* Com-39, 62–75, 1991.

7 Silvester, J, Lee, I. Performance modelling of buffered CSMA : an iterative approach. In: *Proc. GLOBECOM '82,* Miami, FL, 1195–1199, 1982.

8 Takagi, H. *Queueing analysis, vol. 1.* Amsterdam, North-Holland, 1991.

9 Kleinrock, L. *Queueing systems, vol. I.* New York, John Wiley & Sons, 1975.

10 Basic reference model, CCITT/X.200.

11 Service definitions, CCITT/X.210.

Parameterized access modes in Apotram

BY OLE JØRGEN ANFINDSEN

Cooperation between database transactions is desirable in many application domains. This paper presents an approach to this problem based on parameterized read and write operations, a central concept in Apotram – an Application-oriented Transaction Model. Parameterized access modes may be thought of as a generalisation of the well known concept of isolation levels, and enables dynamic cooperation between database transactions without requiring interaction patterns to be predefined. Uncommitted data is considered to be a special case of unreliable data. The quality of uncommitted data is represented by means of reliability indicators. The latter, possibly combined with logic, is used as a language for communication between cooperating transactions.

This is an extended and modified version of a paper that appeared in the proceedings of *2nd International Conference on Concurrent Engineering: Research and Applications,* McLean, Virginia, USA, August 1995.

1 Introduction and motivation

Almost any kind of application could, in principle, benefit from using a database management system (DBMS) for its information storage and retrieval. Advantages of DBMSs are well known, and include data independence, data sharing, concurrent access to data in a controlled fashion, recoverability, and more [10], [16]. The transaction management mechanisms of traditional DBMSs were developed to support short, atomic transactions, e.g. for banking and airline reservation applications. However, some application domains have needs that are different. Frequently mentioned examples include CAD and CASE [15], [41]. A major problem is that such applications need support for long-lived transactions (LLTs) that can handle interaction patterns not allowed by ordinary schedulers. In particular, serialisability (which requires a schedule of concurrent transactions to be equivalent to some serial schedule) is too restrictive a correctness criterion in an LLT environment [32].

Standard transaction management uses flat transactions with ACID properties [26] (ACID stands for Atomicity, Consistency, Isolation, and Durability). There is a close connection between the Isolation part of ACID and the concept of serialisability. A transaction schedule is said to be serialisable if the database is left in a state identical to one that would have resulted were the transactions in question executed in some serial order. One consequence of this is that a database management system (DBMS) that can guarantee serialisable schedules, gives the transactions the illusion that they have the database to themselves, i.e. they appear to execute in isolation. This is a very powerful concept, and is clearly a major factor behind the success of DBMSs. However, there is a price to pay for serialisability and for some application domains it is undesirable [15].

There are basically two different motivations for compromising serialisability: (1) because one is unwilling to pay the cost in terms of resource consumption and/or delays (such as memory consumption, CPU consumption, and time spent waiting for locks), or (2) because the semantics are unsuitable for the application domain in question. In order to deal with the first case, commercially available DBMSs typically offer isolation levels that permit various degrees of non-serialisability. In [2] we argue that our ideas for inter-transaction communication would be useful where serialisability is compromised in the interest of performance, in particular that they would represent a vast improvement over a simple *dirty read* approach. However, this paper will focus on the other case, where non-serialisable schedules is not only something one is willing to accept for pragmatic reasons but something that is inherently desirable or even necessary. Examples will be taken from the domain of design applications. It is well known that such applications need support for cooperative transactions [40], i.e. they need transaction mechanisms that will allow access to uncommitted data in a controlled fashion.

Our approach to transaction cooperation is based on the introduction of parameterized *access modes*, used to specify levels of *protection* and levels of *sharing,* and the use of some language to enable transactions to communicate to each other information about the reliability of uncommitted data. A protection level of a transaction determines its read mode, and a share level determines its write mode. These are central elements in Apotram – an Application-oriented Transaction Model [4].

The remainder of this paper is organized as follows. Section 2 discusses related work. The isolation level concept is introduced in Section 3 and generalized in Section 4. In Section 5 it is shown how generalized isolation levels can be used to support cooperative transactions. Section 6 contains our conclusions.

2 Related work

[19] discusses how to relax the serialisability requirement in order to improve performance. "The main idea is to allow non-serialisable schedules which preserve consistency and which are acceptable to the system users". Such schedules are called Semantically Consistent Schedules. The basic idea is to divide transaction types into disjoint classes. If two transaction types belong to the same class they are *compatible,* i.e. they may be interleaved arbitrarily. Transaction types of different classes are *incompatible,* i.e. they may not be interleaved at all.

This approach is refined by [34] which define hierarchically structured levels of compatibility among transaction types. Transactions should be large in the interest of expressiveness, but divided into several small atomic units which form *breakpoints* in order to allow interleavings. The correctness criterion that replaces serialisability is called multilevel atomicity.

[17] generalizes the approaches of the just mentioned references. Like [34] breakpoints are used to allow levels that fall between arbitrary interleaving and no interleaving, but one no longer requires interleavings to be hierarchically structured. The notion of Relatively Consistent Schedules is defined.

Several papers have been published that address the needs of CAD, CASE, or other engineering applications. [33] introduces a new transaction model called Flex, which allows flexible control of the transaction isolation granularity, distinguishing between compensatable and non-compensatable subtransactions. [24] introduces Dynamic Validation, an optimistic concurrency control scheme that avoids holding locks for long periods of time. [31] defines a model where users are able to influence the setting and releasing of locks by means of check-in and check-out operations, thus enabling cooperation within groups of transactions. Other related papers include [12], [13], [29], [40],

Table 1 Isolation level support in some commercially available DBMSs. Asterisks indicate support. Oracle uses multiversion concurrency control, but their two concurrency control options roughly correspond to QC and TC, respectively, indicated by ◊

Isolation level	DB2	Illustra	Informix	Ingres	Oracle	Rdb	Sybase	UniSQL
UR	*	*	*	*			*	*
CR		*	*			*	*	*
CS	*		*					*
QC					◊			
RR		*				*	*	
TC	*	*	*	*	◊	*	*	*

[41], [42], [44]. The following two references are of particular interest.

[43] generalizes the two-phase locking protocol by introducing altruistic locking which, under certain conditions, enables LLTs to release locks early. This is somewhat similar to our idea of a transactions downgrading its read or write modes to give other transactions access to its data. However, an important difference is that altruistic locking guarantees serialisability.

[30] introduces Database Conversations, "an application-independent, tight framework for jointly modifying common data". In their model, each data unit has a binary *conversation flag* associated with it and the two make up an indivisible unit; "the information whether a data unit is uncertain or not, is stored explicitly with the data item itself and not implicitly in some transaction semantics". Conversation flags correspond somewhat to our *reliability indicators* which also form indivisible units with data attributes: Both are used to tell other transactions something about the status of a data attribute, both are examples of the use of meta-data, and both are based on the notion of uncommitted data being uncertain. But while conversation flags have only two values, reliability indicators can have arbitrarily many values. Thus, this paper shares with (ibid) the basic idea that uncommitted data is unreliable and that transactions should communicate with each other about the status of such data. On the other hand, their language for inter-transaction communication is based on dynamically created *conversation contexts*, while ours is based on reliability indicators combined with multivalued logic.

[20], [21] defines a transaction specification and management environment (TSME), which may be thought of as a programmable scheduler that can be tailored to various needs. The TSME is specifically aimed at object-oriented contexts. [2] as well as this paper have been inspired by the TSME ideas; we attempt to achieve a subset of the TSME goals by means of techniques that are easier to implement and use.

3 Isolation levels

A common way to specify concurrency properties of transactions in DBMSs is isolation levels. This concept is related to our approach to cooperative transactions. Isolation levels were first introduced in [25] under the name *degrees of consistency*. See [26] for a textbook treatment, or [3] for a tutorial style treatment. Although [25] introduced isolation levels in a locking context, they may be considered implementation independent and defined without reference to locking or any other mechanism by which the desired concurrency properties are achieved. This is the case in the SQL92 standard [35] which defines the different isolation levels only by means of various phenomena that they prohibit (for a more readable treatment of SQL92, see e.g. [11], [36]). Traditional isolation levels include:

- *Uncommitted read* (UR, also known as read uncommitted) allows an application to read both committed and uncommitted data.
- *Committed read* (CR, also known as read committed) allows an application to read committed data only.
- *Cursor stability* (CS) allows an application to read committed data only and guarantees that an object will not change as long as a cursor is positioned on it.
- *Repeatable read* (RR) allows an application to read committed data only and guarantees that read data will not change until the transaction commits or aborts (thus, a read that is repeated will return the original row unchanged). RR will not prevent the so called phantom phenomenon, i.e., when a cursor is reopened an object not present the previous time may appear.
- *Transaction consistency* (TC, also known as serialisable) guarantees that the transaction has a consistent view of the database (as if no other transactions were active). We prefer to refer to this isolation level as TC rather than SR (serialisable) because (1) a TC transaction may be part of a non-serialisable schedule where other transactions use lower levels of isolation, and (2) because TC is a more intuitively appealing name in a nested transactions environment. See [2] for a discussion of dynamic usage of isolation levels in a nested transaction context.

Except for CS, the other isolation levels are defined in the SQL92 standard. In (ibid), an additional isolation level called *Query Consistency* (QC), which falls between CS and RR, is defined. QC allows an application to read committed data only, and guarantees that all objects accessed in a single query is consistent, i.e. all objects accessed by one query is from a single, committed state of the database. Implementing QC by means of locking is straightforward; all read locks must be kept until the query is completed (until the cursor is closed).

Note that traditional isolation levels do not in any way influence the handling of write-write or write-read dependencies, only read-write dependencies. This can be seen e.g. by analysing the locking protocols used to implement various isolation levels: The protocols for write locks are always two-phase in that no write locks are released before the transaction terminates (i.e. commits or aborts), irrespective of isolation level. The protocols for read locks on the other hand, are directly influenced by the isolation level. Only with TC (and certain implementations of RR, like e.g. in the DB2 Family which actually provides full TC support under the name RR) are read lock protocols two-phase. With less restrictive isolation levels read locks may be released *before* the transaction terminates, and new ones may still be acquired. Simply put, traditional isolation levels, when implemented by means of locking, determine the *duration* of read locks, and nothing else.

Below we generalize the isolation level concept in such a way that the handling of write-read dependencies is influenced too. This is achieved through the introduction of new read and write modes, and by decomposing isolation levels into *share levels* and *protection levels*.

4 Generalising the isolation level concept

The isolation level concept was introduced in the context of locking, and as far as we know all systems that support isolation levels implement them by means of locking. Although we will try to distinguish clearly between the conceptual and implementation aspects, our treatment of isolation levels will be somewhat biased by a desire to relate our ideas to the locking approach.

4.1 The access mode concept

An important concept when dealing with concurrent use of databases, is that of *access mode*. There are two basic access modes; read and write. When a database object is accessed in read mode, the agent in question can perform only read operations on that object. Two or more transactions may access a given object concurrently, provided they all use read mode. When a database object is accessed in write mode, the transaction in question can perform both read and write operations on that object. More specifically, write mode enables reading, deleting, inserting, and modifying objects. If an object is accessed in write mode by one transaction, no other transaction can access that object in either read or write mode.

In addition to these two basic access modes, many DBMSs support the following three: browse, upgrade, and exclusive. Browse mode enables a transaction to read an object even if some other transaction is currently accessing it in write mode. Thus, when using browse mode, transactions have no guarantee of a consistent view of the database, since there is a risk that they will read uncommitted data. The use of browse mode is often denoted *read through* or *dirty read*, and is used with isolation level UR. Upgrade mode is like read mode with the added semantics that the transaction in question may at any time request an upgrade to write mode, i.e. it may upgrade its access mode. If an object is accessed in upgrade mode by one transaction, other transactions may access that object in read mode, but no other transaction can access that object in either upgrade or write mode. The motivation for upgrade mode is briefly outlined below. If an object is accessed in exclusive mode by one transaction, no other transaction may access that object, irrespective of access mode. Exclusive mode is necessary when even browsers must be kept away, e.g. when a table is to be dropped from a relational database.

Support for upgrade mode was added to DBMSs to prevent a special kind of deadlock, as will now be explained. Some applications work as follows: a number of database objects are "looked at", but only some of these are updated or deleted (the standard example is that of an updatable SQL-cursor, where statements of the type UPDATE WHERE CURRENT OF or DELETE WHERE CURRENT OF may be used). If all the objects in question are "looked at" in write mode, the problem is unacceptably low concurrency. The alternative (assuming upgrade mode is not supported) is to "look at" the objects in read mode and then promote from read to write mode whenever an update or delete operation is to be performed. The problem with this approach is that two transactions may access the same object in read mode, and if they both request promotion to write mode, the result is deadlock. This dilemma is eliminated by supporting upgrade mode, since upgrade modes are not mutually compatible.

Thus, there are five traditional access modes. We will refer to these five access modes frequently in the following, and will therefore denote each of them by a single letter:

B – browse
R – read
U – upgrade
W– write
X – exclusive

We will refer to B, R, and U as read modes, and to W and X as write modes. Later, we will introduce new read and write modes. The relationship between the five access modes described above is nicely captured by the following compatibility matrix:

	B	R	U	W	X
B	*	*	*	*	
R	*	*	*		
U	*	*			
W	*				
X					

Asterisks indicate compatibility. For example, R is compatible with B, R, and U; W is compatible only with B; and X is not compatible with any other access mode.

4.2 Implementation of access modes

Implementing the five access modes by means of locks is fairly straightforward. However, one has to deal with the need to support different lock granularities [25]. Typical granularities could be tuple, object, page, table, class, file, tablespace, or database. Loosely speaking, the granularities of choice will correspond to the resource hierarchy at hand. If a given transaction is to use page locks, say, it will need three page locks with the following lock compatibility matrix:

	R	U	W
R	*	*	
U	*		
W			

As can be seen, there is a one-to-one correspondence between this matrix and a *subset* of the access mode compatibility matrix (it will soon be explained why a subset suffices). There is, by the way, not a universally agreed upon convention for lock names; R locks are often called S (for share), and W locks are often called X (for exclusive). We prefer to denote read locks by R instead of S, since we later introduce locks that are more shareable than these. And since write locks are not truly exclusive, we think W is better than X (some people find it amusing that certain vendors have to refer to the truly exclusive lock as e.g. *super*-exclusive). Upgrade locks are sometimes referred to as update locks, but seem to always be denoted by U.

Aside: In [26] it is suggested that the above matrix should be made asymmetric: If transaction T1 holds an R lock on something, then transaction T2 should be allowed to acquire a U lock on that same thing. However, if T1 holds a U lock, then T2 should not be allowed to acquire an R lock. The advantage is that, in the latter case, T2 will not be able to delay a lock promotion from U to W should T1 request it (such a promotion will of course have to wait until all currently held R locks are released). The disadvantage is that T2 will have to wait for its lock, even if T1 never requests a lock promotion. In (ibid) we define *priority locking* which provides a simple solution to this dilemma. Whether one gets one or the other kind of semantics will then depend on the assigned priorities of the transactions involved. We suspect that symmetric lock compatibility matrices are more common in commercially available DBMSs (Informix and the DB2 Family are examples), so that is what will be used in this paper. **End of aside**.

Unless it is required that all transactions use the same lock granularity, one must be able to coordinate concurrent transactions that request locks at different levels in the resource hierarchy. The solution given in [25] and implemented in a large number of commercial systems, is as follows: a transaction will request R, U, and W locks at some level of choice, it will not request any locks of lower levels, but it will request *intent* locks at all higher levels. (Simplifications of this procedure by means of clever tricks with low-level physical mechanisms will not be considered here.) It is possible for a transaction to use different lock granularities for different statements, but this is not significant for the discussion at hand. Two basic intent locks are needed; IR indicates an intent to request R locks at some lower level, and IW indicates and intent to request W locks at some lower level. No IU lock is needed, because a transaction that intends to request U locks at some lower level also has an implicit intent to request W locks (otherwise there would be no need for U locks in the first place). Thus, such a transaction must use IW. There turns out to be a need for one more lock, RIW, which is a combination of R and IW; it provides R-access to the entire resource in question (e.g. a class extension) while also enabling the transaction to request U and W locks at some lower level (e.g. page or object). The complete lock compatibility matrix looks like this:

	B	IR	R	U	IW	RIW	W	X
B	*	*	*	*	*	*	*	
IR	*	*	*	*	*	*		
R	*	*	*	*				
U	*	*	*					
IW	*	*			*			
RIW	*	*						
W	*							
X								

The necessity of an exclusive access mode (and hence an X lock to implement it) was briefly outlined above. Note that the RIW row/column is identical to the intersection of the R and IW rows/columns. In practice, B locks, and hence X locks, are not used at the lowest levels (the plural form is used here, since using B locks at the lowest levels would – partially at least – have defeated the purpose of using isolation level UR in the first place). For example, in a relational DBMS a UR transaction will typically request a B lock at the table level (and all levels above), and then proceed without requesting any locks on pages or tuples. However, it will request some sort of low-cost, short duration locks (known as latches) on pages or tuples to ensure atomicity of individual read operations.

In order to figure out which locks are compatible with which, simply consider the definitions of the access modes involved. For example, the R lock is compatible with the other read locks (B, IR, and U), but not compatible with any of the write locks (IW, RIW, W, and X). Or consider the IW lock; obviously compatible with B, compatible with IR since potential conflicts between IW- and IR-transactions are resolved at some lower level, not compatible with R and U since these locks preclude writers, compatible with IW since potential conflicts between two IW-transactions are resolved at some lower level, and finally IW is of course not compatible with W and X locks. By reasoning like this the entire matrix follows.

Some vendors denote browse locks with IN, for intent none. Just like denoting read locks with S, this seems to reflect the perspective of implementors rather than users. Since the work presented here is part of a project whose goal is to define and describe Application-Oriented Transaction Management (AOTM), we prefer things to be the other way round, i.e. we think names should be chosen with a user perspective in mind.

We conclude this section with some brief examples (assuming for now an SQL environment). For an RR transaction that wants to scan an entire table, it is probably a good idea (for the DBMS component that makes the decision) to request an R lock on the table in question, and then IR locks on the tablespace and database. If the same transaction used isolation level CS instead of RR, R locks on pages and IR locks on the levels above might be a better idea. If a CS transaction wants to scan (parts of) a table with an updatable cursor, it may use U locks (which will be promoted to W locks as needed) on pages and IW locks at the levels above. If, for some reason, the latter transaction requests an R lock on the table that it already has an IW lock on, that IW lock will be promoted to an RIW lock (IW locks at higher levels remain unaffected). It may be worth noting that (even though the name does not indicate so) an IW lock on a table, say, enables its holder to request R, U, and W locks (not just W locks) on tuples or pages. This is quite natural since W locks are stronger than the other two.

4.3 New access modes

We are now ready to extend the concepts described above. We have seen that UR transactions can read uncommitted data. Thus, reading in browse mode could be thought of as a *very* primitive kind of cooperation between transactions. It is primitive because (1) there is – in general – no knowledge of who is cooperating with whom, (2) browsers do not get to know whether data they read is committed or not, and (3) whenever uncommitted data is encountered, the browsers receive no information about the quality or status of the data. Thus, for applications that need the data sharing kind of cooperation (design and engineering applications are classical examples), using UR transactions is an unsatisfactory solution.

As a first step away from the blind cooperation provided by UR, we introduce *parameterized* read and write modes. The basic idea here is that users should be able to specify when reading and writing should be incompatible. In other words, the standard notion that read and write modes are mutually incompatible is reduced to a *default* which transactions may override by proper use of parameters.

Let D denote the domain of access mode parameters, and A, B ⊆ D. Parameterized read and write modes will be denoted as R(A) and W(B), respectively. R(A) and W(B) are compatible iff B ⊆ A (iff is shorthand notation for if and only if). For example, if u1, u2, u3 ∈ D then it would be the case that

- R(u1) and W(u1) are compatible
- R(u1) and W(u2) are incompatible
- R(u1, u2) and W(u2) are compatible
- R(u2) and W(u2, u3) are incompatible.

By means of generalized isolation levels, users may specify the parameters that will be used during read and write access. Non-parameterized read and write modes can still be denoted as R and W but should, in the interest of generality, be thought of as R(Ø) and W(*), respectively, where Ø denotes the empty set and D ⊂ * (i.e. * denotes an arbitrary superset of D). Thus, according to the rule that R(A) and W(B) are compatible iff B ⊆ A, R(Ø) will be incompatible with all write modes and W(*) will be incompatible with all read modes. Conversely, R(D) denotes the read mode that is compatible with all parameterized write modes (but not with W(*)).

A parameterized write mode indicates willingness to share information with readers. The idea is that a parameterized writer indicates to other transactions that it is *relatively safe* to read its data. This could be so simply because statistical analysis shows that some transactions hardly ever abort. Or in e.g. a design environment one may reach a point in the process where the design has stabilized sufficiently that other transactions should be given read access. Analogously, the use of parameterized read modes indicates willingness to read data that belongs to parameterized writers.

The new access mode compatibility matrix is a trivial extension of the old one:

	B	R(A)	U	W(B)	X
B	*	*	*	*	
R(A)	*	*	*	?	
U	*	*			
W(B)	*	?			
X					

Compatibility matrix for access modes: * indicates compatibility and ? indicates compatibility iff B ⊆ A.

4.3.1 Remarks

The use of parameterized access modes may be thought of as associating *conditions* with the access modes; conditions that must be evaluated in order to determine compatibility. One could imagine read modes that have more general conditions associated with them. Such conditions could be implicitly given by the query performing the read operations (i.e. by the predicates of the query), or explicitly given by extra predicates. Conditional access modes are not important concerning the understanding of our approach to cooperative transactions. However, they are briefly mentioned here partly to indicate the versatility of this approach, and partly to answer questions like the following: What if a transaction is willing (or even wants) to read uncommitted data of a certain quality, but still needs to scan data more than once, how can such requirements be combined? That is, if e.g. a transaction with isolation level repeatable read scans part of the database in read mode R(A), A ≠ Ø, for the second time, could it not be the case that the data in question could have been deleted since the first scan, or modified in such a way that it no longer satisfied the search criteria? And yes, the latter scenario is possible unless precautions are taken. One approach could be to put restrictions on the allowable locking protocols, such as e.g. prohibiting deletions to be performed with W(B), B ≠ *, locks. Probably a good idea in and of itself, but unfortunately it solves only part of this problem. Another approach could be to provide readers with needs of the kind mentioned above with conditional read modes. Conditional access modes as well as the locks needed to implement them are discussed in [2].

4.4 Implementation of new access modes

Parameterized access modes can very naturally be mapped to corresponding parameterized lock modes (assuming all the time that locking is the chosen implementation method). The only tricky thing is that RIW needs both a read and a write parameter (in general), and is thus denoted R(A)IW(B). The semantics of this lock can be paraphrased as read with parameter set A and intention to write with parameter set B. Two locks R(A1)IW(B1) and R(A2)IW(B2) will be mutually compatible iff B1 ⊆ A2 and B2 ⊆ A1. The lock compatibility matrix now becomes:

	B	IR(A)	R(A)	U	IW(B)	R(A)IW(B)	W(B)	X
B	*	*	*	*	*	*	*	
IR(A)	*	*	*	*	*	*	?	
R(A)	*	*	*	*	?	?	?	
U	*	*	*					
IW(B)	*	*	?		*	?		
R(A)IW(B)	*	*	?		?	??		
W(B)	*	?	?					
X								

Compatibility matrix for lock modes: * indicates compatibility, ? indicates compatibility iff B ⊆ A, and ?? indicates compatibility iff B ⊆ A both ways (i.e. from lock holder to lock requester, and vice versa).

We do not consider parameterized lock modes a serious runtime overhead problem. Held locks reside in (more or less) dynamically allocated storage blocks, each the size of several bytes (identifying the held resource, the holding transaction, lock duration, lock mode, and more). Space requirements per lock need not grow more than about one bit per member of the parameter domain D.

4.4.1 Remarks

A vendor that were to implement generalized isolation levels would probably want to give careful consideration to details of locking protocols. For example, one may want to temporarily upgrade from W(B) to W while a modification of a locked object is performed, and then downgrade afterwards, that way parameterized readers will not experience that data objects change while they are positioned on them (this problem is not solved by the latches mentioned above). A discussion of locking protocols is considered beyond the scope of this paper.

Likewise, this paper does not allow for a discussion of recovery considerations. However, traditional recovery schemes can be used with generalized isolation levels. Support for this claim will be given in [4].

4.5 Parameterized access modes and isolation levels

As already mentioned, traditional isolation levels deal with read-write dependencies only, and in terms of locking this means varying the *duration* of read locks. The introduction of parameterized read modes modifies the way read-write dependencies are handled. Combining parameterized read modes with isolation levels should be no problem, it simply means varying the duration of R(A) locks rather than plain R locks.

Isolation levels have no influence on write lock durations; write locks are never released until termination, or else one would not be able to guarantee recoverability [6].

5 Cooperative transactions

At the very core of our approach to cooperative transactions is the following observation: uncommitted data contains unreliable information, i.e. it is associated with some level of uncertainty. Thus, dealing with uncommitted data is just a special case of the general problem of dealing with uncertainty. Many papers have been published on this topic. [14] contains “an evolving bibliography of documents on uncertainty and imprecision in information systems” (but apparently unchanged since 1993), and currently gives 352 references. See e.g. [37], [38] for surveys.

According to [38] “most approaches to the modelling of imprecision in databases are based on the theory of fuzzy sets”. An alternative approach is to use probability distributions (rather than the possibility distributions found in fuzzy set theory). Examples of the latter can be found in e.g. [5], [18], [28]. Both in fuzzy and probabilistic approaches, users will have to associate with data attributes numbers in the interval (0, 1) to denote degrees of membership or probabilities, respectively. Choosing numbers that faithfully reflect reality is clearly a non-trivial task, and [5] acknowledges that this “could be a problem in practice because it may be difficult to know the exact probabilities”.

We have therefore chosen a different approach. Instead of forcing users to *quantify* uncertainty, we allow them to *classify* it. That is, users can denote unreliable data as belonging to one or more of a predefined set of unreliability indicators. This set must be chosen so as to meet the needs of the application. However, we would like to emphasize the following: Our approach to transaction cooperation is not dependent on a particular approach to uncertainty management. We are simply using the approach to uncertainty management that appears to be most suitable for the task at hand. Likewise, a decision to use unreliability indicators rather than fuzzy or probabilistic approaches, does *not* force one to use a particular logical framework such as X5VL.

Before we go into the details of the cooperation mechanisms, we describe an example scenario that illustrates the purpose of those mechanisms. The scenario is not intended to be realistic, but should give a rough idea of the problems we are trying to address.

5.1 Example scenario

Consider an engineer working on the design of some part of a new aeroplane, e.g. the landing gear, using a CAD system with transactional capabilities. Initially, the design is constantly changing. The engineer probably has a lot of information about the new landing gear she is about to design; she might know that her company has signed a contract with a new supplier of tyres, that a decision has been made to use a new hydraulic technology, etc. Whether she copies the landing gear of an existing aeroplane into her work space or she chooses to work from scratch, there will be a period of time when her design is unstable and inconsistent. During this period it will make sense to protect her design by ordinary write locks, and her CAD system could automatically put the very lowest reliability rating on the design as a whole.

Sooner or later the design reaches a stage where it can be considered a draft; it is complete, it is consistent, and it meets overall design criteria that were given in advance (having to do, say, with weight and physical dimensions). This is when the need to cooperate with other designers come into full play. She will need to look at the design of other parts of the aeroplane, such as the wings, the fuel tanks, the hydraulic system, the electric system, etc. Conversely, other designers will need to look at the landing gear design. A number of modifications to many parts of the aeroplane is to be expected, many of which will make other modifications necessary. For example, it might be necessary to alter the physical placement of some hydraulic pipes, forcing our engineer to modify the landing gear design accordingly.

Designing an entire aeroplane is an enormous task, and the work evolves through several stages. Presumably, there must be a company policy in place that specifies various levels of approval that designs can achieve. And e.g. the designer of the overall hydraulic system would be interested in knowing the level of approval achieved by designs that involve hydraulic components. We suggest that these levels of approval that a given design goes through during its lifetime, should be reflected in the reliability indicators that we propose. In other words, we think it is a good idea that the CAD system be able to represent levels of approval.

Assuming that the many engineers involved in designing an aeroplane are not always synchronized, i.e., that their designs may be at different levels of approval, one can easily conceive of managers and others wanting to produce reports that help them assess progress. Thus, we think it is a good idea to have a connection between reliability indicators and the query language supported by the system.

5.2 Reliability indicators

As mentioned above, our approach is based on regarding uncommitted data as unreliable, and we choose to represent the latter by means of reliability indicators. Two important things should be noticed at this point. First, our approach does not say anything about the *semantics* of reliability indicators. Second, our approach does not say anything about the *number* of different reliability indicators. This means that in one database one may define one set of reliability indicators, and assign to each of them a specific meaning. In another database one may define another set of reliability indicators, possibly with completely different semantics. Thus, this approach could be tailored to many different application domains.

Without going into implementation details (which we consider irrelevant here), we are assuming that reliability indicator definitions, along with a description of their semantics, are stored in a meta-database to which users have read access. It is assumed that interpretation of reliability indicator semantics takes place at the application level. In other words, an application or its user must know how to interpret reliability indicators when they are encountered, and which reliability indicators to use when updating the database. If users and applications are aware of the way reliability indicators are being used in their system, they have a *language* by which to communicate when running transactions.

5.3 Reliability indicators and logic

In a database that does not deal with uncertainty, the result of a scalar comparison like e.g. Salary > 50000 is either true or false, period. However, if a Salary value is known to be unreliable, then one cannot simply return true or false from the above predicate; very loosely speaking, predicate evaluations become as uncertain as the data involved. Considerations like this give rise to a plethora of multivalued logics. A well known and extreme case is when data is not only unreliable, but completely missing; SQL-style nulls are a well known way of representing missing information, and are supported by many DBMSs, result in three valued logic [7], [9]. [8], [22] advocate the use of two semantically different nulls and corresponding four valued logics. [23] propose a five valued logic. And e.g. the probabilistic approach of [28] gives rise to any number of *degrees of truth* in the interval [0, 1]. Many other logics are defined by papers listed in [14].

The reliability indicator concept has its roots in the null values approach to the problem of missing information [7], [1], and we consider it to be a special case of the ideas advocated in [39]. In order to perform scalar comparisons, evaluate complex predicates, and perform other kinds of manipulations of data values (which may or may not be annotated with reliability indicators) in a well defined manner, we employ a framework called Extended Five Valued Logic (X5VL). We refer to X5VL as a *framework* since one must define some set of reliability indicators (as well as a set of nulls, if missing information must be handled) before one can start using it. Since, as mentioned above, one could use some other logical framework than X5VL in this context, only limited treatment of X5VL is given here. Examples illustrating its use with cooperative transactions are given below, and a very brief overview of X5VL is given in the Appendix. A full description of X5VL is given in [1].

5.4 Example queries

Consider once more the aeroplane designer scenario and a CAD system based on the X5VL framework. Assume that designs evolve through the approval levels a0, a1, ... , a5, where a0 is used in the initial phase and a5 is the final level of approval. Assume therefore that, corresponding to the approval levels, reliability indicators a0 through a5 have been defined in the meta-database. Also, assume that various parts of a design can be at various levels of approval.

The following examples are meant to illustrate some of the expressive power of X5VL. The syntax used is hypothetical, and should not be interpreted as concrete proposals. We take the liberty of deviating somewhat from the syntax constructs defined in [1].

As in the above scenario, the examples are not meant to be quite realistic. We are not making assumptions about the data model of the CAD system (relational or object-oriented), only about the availability of an SQL-like query language.

Example 1:

We want to get an overview of the landing gear design, provided that a certain level of approval has been reached for some key components. We would only want to retrieve information if nuts, bolts, and hydraulics have achieved a4 or higher. The following query, with read mode R(D), may then be used (assuming that DESIGNS identify a set of objects, each of which has attributes NAME, NUTS, BOLTS, and HYDRAULICS):

```
SELECT  <whatever>
FROM    DESIGNS
WHERE   NAME = 'Landing gear'
 AND    QUALITY(NUTS, BOLTS, HYDRAULICS)
        IN (a4, a5)
```

QUALITY is a hypothetical operator that simply extracts the values of the reliability indicators of the attributes given to it as parameters. These values are compared to those of the IN list, yielding true or false. If the CAD system offered no X5VL support, and levels of approval were stored in ordinary attributes (NUTS_APPROVAL and BOLTS_APPROVAL, say) rather than reliability indicators (which may be thought of as meta-attributes), one would need this kind of a query instead:

```
SELECT  <whatever>
FROM    DESIGNS
WHERE   NAME = 'Landing gear'
 AND    NUTS_APPROVAL IN ('a4','a5')
 AND    BOLTS_APPROVAL IN ('a4','a5')
 AND    HYDRAULICS_APPROVAL IN
('a4','a5')
```

The point is that the use of reliability indicators enables the system to offer a lot of expressive power while the more ad-hoc approach forces the user to explicitly identify all possible cases.

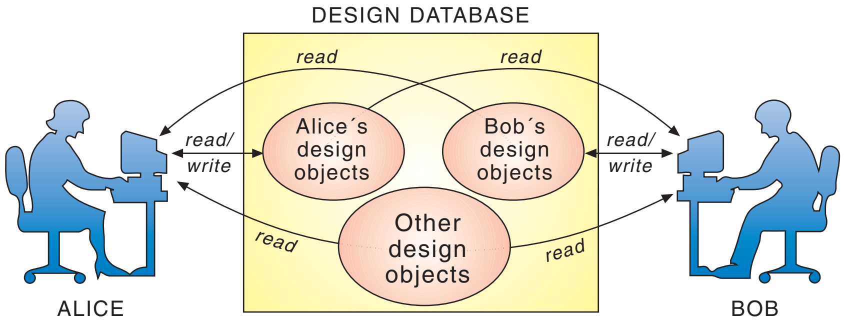


Figure 1 Design applications need support for cooperative transactions. As a minimum, one designer must be able to look at design objects currently being worked on by another designer

The difference becomes greater if the number of attributes or reliability indicators is increased.

Example 2:
We would like information about all designs of things that occupy little space, but weigh and cost a lot. Again, the reliability level must be a4 or higher, and the attributes named below are assumed to be present in design objects. The following query, with read mode R(D), may be used:

```
SELECT  <whatever>
FROM    DESIGNS
WHERE   TRUTHVALUE (SPACE < 5
 AND    WEIGHT > 777
 AND    COST > 999)
        IN QTRUE(a4, a5)
```

The important thing here is the use of the hypothetical TRUTHVALUE operator, which takes as its parameter an arbitrarily complex predicate, followed by the keyword IN and then a truth value specification (soon to be explained). In an X5VL environment predicates that involve annotated values do not evaluate to true or false, instead they evaluate to a member of the set qtrue (quasi-true) or qfalse (quasi-false), respectively. For example, if SPACE contains the value 3 annotated with reliability indicator a4, the predicate SPACE < 5 evaluates to the a4-member of the qtrue set. However, SPACE > 5 would evaluate to the a4-member of the qfalse set. In other words, qtrue values are true-like but weaker than true, and qfalse values are false-like but weaker than false. X5VL defines the result of any combination of true, qtrue, qfalse, and false truth values (as well as a fifth class referred to as maybe (X5VL without the maybe class is X4VL)), and the VERIFY clause is used to verify that the predicate before IN evaluates to the range of truth values specified after IN. According to the rules of X5VL, if SPACE, WEIGHT, and COST are all annotated with a4 or a5, the above predicate will evaluate to a4, a5, or a4a5 in the qtrue class, i.e. QTRUE(a4, a5). If at least one of SPACE, WEIGHT, and COST is annotated with something else than a4 or a5, the above predicate will evaluate to something outside of QTRUE(a4, a5).

If the CAD system supported cooperative transactions and reliability indicators, but had not integrated the latter with X5VL or some other logic, one would have less expressive power. The above query would have to be written as:

```
SELECT      <whatever>
FROM  DESIGNS
WHERE SPACE < 5
 AND  WEIGHT > 777
 AND  COST > 999
 AND  SPACE_APPROVAL IN (‘a4’,‘a5’)
 AND  WEIGHT_APPROVAL IN
(‘a4’,‘a5’)
 AND  COST_APPROVAL IN (‘a4’,‘a5’)
```

Again, the difference between the two approaches would have been more dramatic had the query involved more attributes and/or reliability indicators.

5.5 General remarks

It is possible for a piece of data to simultaneously be unreliable in more than one way. For example, a geographical information system may store satellite photos that have different resolutions, were taken at different angles, and have different ages; these are three dimensions of unreliability. A piece of data may be inaccurate because it is an estimate, but if it has been written by a not yet committed transaction, that adds another dimension to its unreliability. Adding features to X5VL to support multidimensional unreliability is straightforward [1]. Thus, X5VL may be used not only as a language for communication between cooperating transactions, but also to integrate this with the ability to deal with missing and (several kinds of) unreliable information in general.

We believe the generality of our approach is an advantage in several ways. Consider e.g. this pragmatic argument. The DBMS vendors’ primary concern is to make money, therefore their development is driven by market demand. Because the engineering community is only a small fraction of the DBMS market, their needs have traditionally not been terribly successful in the competition for the *n* topmost places on vendors’ lists of features to be added in the next release of their products, at least this seems to have been the case with the major relational DBMS vendors. Thus, we believe it is an advantage that the ideas presented in this paper are not limited to use in engineering applications only, but have a number of potential uses in business applications as well [1], [2]. In the latter reference we discuss the possibility of a single transaction supporting multiple isolation levels, i.e. various parts of a transaction could be executed with different isolation levels. We argue that a better alternative than assigning isolation levels to individual queries, is to support a nested transaction model and allow specification of isolation levels for each subtransaction. We are still pleased to see in [27] that Version 4 of DB2 for MVS will support specification of isolation levels for individual queries (no member of the DB2 Family supports a nested transaction model yet), and see this as a confirmation of the usefulness of Application-Oriented Transaction Management (AOTM).

6 Conclusions

This approach to cooperative transactions and data sharing generalizes a well understood and widely used concept, isolation levels. Therefore, we believe it would be relatively simple to implement and use. The use of X5VL (or some other logic) as a language for inter-transaction communication enables co-operation to be dynamic, rather than having to be predefined and static. It is an advantage that our approach can be integrated with mechanisms for dealing with unreliable and missing information in general, and that both the X5VL- and AOTM-concepts are relevant in a number of application domains.

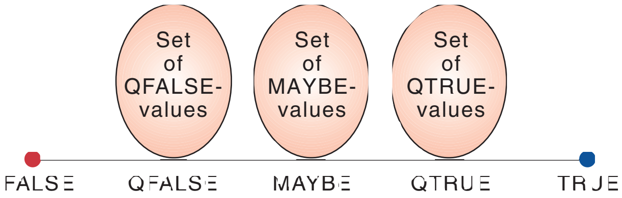


Figure 2 The truth values of X5VL

Acknowledgements

The author wishes to thank Mark Hornick and Dimitrios Georgakopoulos for numerous valuable discussions during my one year stay as a guest researcher at GTE Laboratories.

Appendix A

Extended Five Valued Logic (X5VL) [1] is a framework for dealing with reliable, unreliable, and missing data. It allows *meta-information* to be attached to data, to indicate its state (reliable, unreliable, missing) and to specify *how* data is unreliable or *why* it is missing. X5VL defines *default* results for expressions involving unreliable or missing information, however, user-defined functions may override these defaults. X5VL also defines results of truth-valued expressions and an algebra of truth values.

In addition to **true** and **false** there are a finite number of **maybe** values, corresponding to nulls and combinations of nulls; there are also a finite number of **qtrue** (quasi-true) and **qfalse** (quasi-false) values, corresponding to unreliability categories and combinations of these. The truth values of X5VL are illustrated in Figure 6.1. X5VL is radically different from fuzzy logic and probabilistic approaches to uncertainty management since no *quantification* of uncertainty is made.

When using X5VL as a tool with cooperative transactions, we are *not* interested in nulls and/or **maybe** values, but in unreliable information and quasi-true and quasi-false truth values. Data “owned” by an updating transaction is unreliable because it is subject to change, hence, transactions could benefit from distinguishing between different kinds of unreliability in such a context. Earlier in this paper we discussed annotating data with levels of approval. In more general cases uncommitted data could be labelled as *candidate for deletion, candidate for update* (or more specifically *candidate for increase* or *candidate for decrease*), *in risk of rollback*, *likely to be committed*, etc. Different unreliability categories are appropriate in different application domains. The X5VL framework enables manipulation of and reasoning about *any number* of unreliability categories (subject only to implementation limits). Moreover, X5VL has well-defined and intuitively appealing behaviour in the presence of *any combination* of ordinary (reliable) data, unreliable data, and nulls.

A major advantage with X5VL is that it provides a *uniform and integrated* framework for these aspects of data manipulation. An enhanced version of X5VL, called Annotated Three Valued Logic (A3VL), will be presented in [4].

References

1 Anfindsen, O J, Georgakopoulos, D, Normann, R. *Extended three and five value logics for reasoning in the presence of missing and unreliable information.* Kjeller, Norwegian Telecom Research, 1994. (Technical report TF R 9/94.) (http://www.nta.no/brukere/olea/pub/).

2 Anfindsen, O J, Hornick, M. *Application-oriented transaction management.* Kjeller, Norwegian Telecom Research, 1994. (Technical report TF R 24/94.) (http://www.nta.no/brukere/olea/pub/).

3 Anfindsen, O J, Hornick, M. Isolation levels in relational database management systems, *Telektronikk,* 90, (4), 71–72, 1994. (http://www.nta.no/brukere/olea/pub).

4 Anfindsen, O J. *Apotram : an application-oriented transaction model* (preliminary title). PhD thesis, UNIK center for technology at Kjeller, Department of Informatics, University of Oslo, 1996. Forthcoming.

5 Barbara, D, Garcia-Molina, H, Porter, D. A probabilistic relational data model. In: *Proceedings of the international conference on extending database technology : advances in database technology (EDBT’90).* Venice, 1990, 60–74.

6 Bernstein, P A, Hadzilacos, V, Goodman, N. *Concurrency control and recovery in database systems.* Reading, Mass., Addison-Wesley, 1987.

7 Codd, E F. Extending the database relational model to capture more meaning. *ACM transactions on database systems,* 4, 397–434, 1979.

8 Codd, E F. *The relational model for database management : version 2.* Reading, Mass., Addison-Wesley, 1990.

9 Date, C J. Null values in database management. In: *Proceedings of 2nd British national conference on databases.* Bristol, 1982, 147–166.

10 Date, C J. *An introduction to database systems, volume 1* (fifth edition). Reading, Mass., Addison-Wesley, 1990.

11 Date, C J, Darwen, H. *A guide to the SQL standard.* Reading, Mass., Addison-Wesley, 1993.

12 Dayal, U, Hsu, M, Ladin, R. Organizing long-running activities with triggers and transactions. In: *Proceedings of ACM SIGMOD.* Atlantic City, N.J., 1990, 204–214.

13 Dayal, U, Hsu, M, Ladin, R. A transactional model for long-running activities. In: *Proceedings of the 17th VLDB conference.* Barcelona, 1991, 113–122.

14 Dyreson, C E. *A bibliography on uncertainty management in information systems.* Department of computer science, University of Arizona, 1993. (ftp from cs.arizona.edu).

15 Elmagarmid, A K (ed.). *Database transaction models for advanced applications.* San Mateo, Calif., Morgan Kaufmann, 1992.

16 Elmasri, R, Navathe, S B. *Fundamentals of database systems.* Redwood City, Calif., Benjamin/Cummings, 1989.

17 Farrag, A A, Ozsu, M T. Using semantic knowledge of transactions to increase concurrency. *ACM transactions on database systems,* 14, 503–525, 1989.

18 Feng, Z, Jia, Y, Miller, M. Null values in relational DBMS. In: *Proceedings of the 2nd Australian database & information systems conference.* 1991, 76–86.

19 Garcia-Molina, H. Using semantic knowledge for transaction processing in a distributed database. *ACM transactions on database systems,* 8, 186–213, 1983.

20 Georgakopoulos, D, Hornick, M, Krychniak, P. An environment for the specification and management of extended transactions in DOMS. In: *Proceedings of the third international workshop on RIDE-IMS'93,* Vienna, 1993, 253–257.

21 Georgakopoulos, D et al. Specification and management of extended transactions in a programmable transaction environment. In: *Proceedings of the 10th international conference on data engineering,* Houston, Texas, 1994, 462–473.

22 Gessert, G H. Four valued logic for relational database systems. *ACM SIGMOD record,* 19, (1), 29–35, 1990.

23 Gottlob, G, Zicari, R. Closed world databases opened through null values. In: *Proceedings of the 14th VLDB conference.* Los Angeles, California, 1988, 50–61.

24 Goyal, P, Narayanan, T S, Sadri, F. Concurrency control for object bases. *Information systems,* 18, (3), 167–180, 1993.

25 Gray, J N et al. Granularity of locks and degrees of consistency in a shared database. In: *Proceedings of IFIP working conference on modelling of data base management systems.* Freudenstadt, Germany, 1976, 695–723. Also in *Modelling in data base management systems,* G M Nijssen, ed., Elsevier North-Holland, 1976, 365–395.

26 Gray, J, Reuter, A. *Transaction processing : concepts and techniques.* San Mateo, Calif., Morgan Kaufmann, 1993.

27 Hoover, C, Rizner, J, Yevich, R. DB2 version 4 : our dreams come true? *IDUG solutions journal,* 1, (2), 1994, 12–25.

28 Jia, Y, Feng, Z, Miller, M. A multivalued approach to handle nulls in RDB. In: *Future database'92 : proceedings of the second Far-East workshop on future database systems.* Kyoto, Japan, 1992, 71–76.

29 Kim, W et al. A transaction mechanism for engineering design databases. In: *Proceedings of the 10th VLDB conference.* Singapore, 1984, 355–362.

30 Kirsche, T et al. Cooperative problem solving using database conversations. In: *Proceedings of the 10th international conference on data engineering,* Houston, Texas, 1994, 134–143.

31 Klahold, P et al. A transaction model supporting complex applications in integrated information systems. In: *Proceedings of ACM SIGMOD international conference on management of data.* 1985, 388–401.

32 Korth, H F, Speegle, G. Formal aspects of concurrency control in long-duration transaction systems using the NT/PV model. *ACM transactions on database systems,* 19, 492–535, 1994.

33 Leu, Y, Elmagarmid A K, Boudriga, N. Specification and execution of transactions for advanced database applications. *Information systems,* 17, (2), 171–183, 1992.

34 Lynch, N. Multilevel atomicity : a new correctness criterion for database concurrency control. *ACM transactions on database systems,* 8, 484–502, 1983.

35 Melton, J (ed.). *Information processing systems : database language SQL 2.* (ISO/IEC 9075 : 1992.)

36 Melton, J, Simon, A R. *Understanding the new SQL : a complete guide.* San Mateo, Calif., Morgan Kaufmann, 1993.

37 Motro, A. Accommodating imprecision in database systems : issues and solutions. *ACM SIGMOD record,* 19, (4), 69–74, 1990.

38 Motro, A. Imprecision and incompleteness in relational databases : survey. *Information and software technology,* 32, (9), 579–588, 1990.

39 Motro, A. Annotating answers with their properties. *ACM SIGMOD record,* 21, (1), 54–57, 1992.

40 Nodine, M, Ramaswamy, S, Zdonik, S. A cooperative transaction model for design databases. In: *Database transaction models for advanced applications.* Ed. A Elmagarmid. San Mateo, Calif., Morgan Kaufmann, 1992, 53–86.

41 Nodine, M, Zdonik, S. Cooperative transaction hierarchies : transaction support for design applications. *VLDB journal,* 1, (1), 41–80, 1992.

42 Rauft, M A, Rehm, S, Dittrich, K R. How to share work on shared objects in design databases. In: *Proceedings of the international conference on data engineering,* 1990, 575–583.

43 Salem, K, Garcia-Molina, H, Shands, J. Altruistic locking. *ACM transactions on database systems,* 19, (1), 117–165, 1994.

44 Skarra, A H. A model of concurrent cooperating transactions in an object-oriented database. In: *Lecture notes in computer science 492,* 352–368, 1991.

Using Apotram to alleviate problems in distributed database systems

BY OLE JØRGEN ANFINDSEN

Communication and node failures can cause serious problems for distributed database systems (DDBSs). In particular, blocking is an inevitable problem when atomic commit protocols are used. This paper shows that blocking can be a less severe problem with lenient isolation Apotram transactions. In the proposed scheme a blocked transaction can make its locked data available to other readers in a controlled fashion. The lenient isolation properties also make more generous time-out limits as well as elaborate commit, termination, and recovery protocols feasible. This is a foundation for DDBSs that can provide different semantics to different application domains, thereby making true DDBS functionality attractive to distributed computing applications that currently use ad hoc solutions. It is also discussed how DDBS applications can take advantage of lenient isolation Apotram transactions.

1 Introduction

An intrinsic property of distributed database systems (DDBSs) is that they may encounter communication or node failures [18]. Unless an atomic commit protocol (ACP) such as two phase commit (2PC) is used, inconsistent commit/abort decisions can be reached by distributed transaction participants. However, the combination of the just mentioned failures and ACPs can give rise to situations where participants in a distributed transaction are unable to reach a commit/abort decision. Such (sub)transactions will have to wait for an unbounded period of time. This phenomenon is known as *blocking,* and can be a serious problem since resources may be held indefinitely [4].

In order to minimize such problems, co-operative termination protocols (CTPs) may be employed (ibid). Participants in a distributed transaction that are *uncertain* (explained below) can communicate with each other according to some agreed upon CTP hoping to reach a consistent abort/commit decision. A CTP is invoked when a participant waiting for a message from the transaction co-ordinator reaches a time-out limit.

This paper presents results that (1) make blocking less severe, (2) make more elaborate commit, termination, and recovery protocols feasible, and (3) make the use of more generous time-out limits a viable option. Those results are based on Apotram – an Application-oriented Transaction Model [3] – which is capable of compromising isolation in a controlled fashion, while to a large extent retaining the other ACID properties.[1] It is shown in [2] that Apotram has desirable properties for co-operative transactions such as found in e.g. design environments. The goal of this paper is to show that Apotram transactions with lenient isolation properties are useful when dealing with the just mentioned DDBS problems. The reader is assumed to be at least somewhat familiar with the pertinent aspects of the transaction model – presented in Parameterized Access Modes in Apotram elsewhere in this issue of *Telektronikk.*

Part of the philosophical underpinning of Apotram is the pragmatic attitude that compromises should be found when a perfect solution (read serialisability) cannot be attained or has unreasonable consequences. A relevant quote can be found in [5]: “full data consistency and serialisability can only be achieved in a multidatabase system by imposing restrictions that many consider severe. Thus, there is a need to identify alternative forms of consistency and ways of restricting standard notions of consistency so that positive results can be stated, rather than impossibility results.” It is my belief that this principle should be applied not only to multidatabases but to DDBSs and other problem domains as well.

The remainder of this paper is organized as follows. Section 2 briefly mentions some related work. Section 3 shows how Apotram can be employed at the *system level* of a distributed database. Section 4 shows how Apotram can be employed at the *application level* of a distributed database. Section 5 contains my conclusions.

2 Related work

[9], [16], and [8] all discuss how to relax the serialisability requirement in order to improve performance. They introduce the alternative correctness criteria Semantically Consistent Schedules, Multilevel Atomicity, and Relatively Consistent Schedules, respectively. The latter is used in [20] as a foundation for Semipermeable Transactions in multidatabases.

Apotram shares with the Transaction Specification and Management Environment (TSME) [10], [11] the concept that transaction management should be tailored to meet the needs of applications. In fact, it was during my one year stay as a guest researcher at GTE Laboratories in 1993–94, where I was exposed to TSME, that the ideas that lead to Apotram were conceived.

Several papers that have fundamentally different approaches to node or communication failures in DDBSs have been published. One example is [15] that uses *relaxed atomicity* which involves uncoordinated local commit or abort decisions, making it necessary for “transactions with discrepant commit decisions [to be] recovered semantically rather than physically”. Another example is [22] that divides data values into *partitions* which are placed on different nodes. Since access to a single such partition will be sufficient for several classes of applications, this results in increased availability of data and more.

3 Apotram and distributed systems

It is not my intention to carry out an exhaustive discussion of all possible problem scenarios, but rather to show how Apotram could be used in well-known situations described in standard textbooks such as e.g. [4], [18].

3.1 Blocking

A participant in a distributed transaction is said to be *uncertain* if it has responded to the decision request from its coordinator with a Yes vote, but has not yet received a final Abort/Commit message. An uncertain participant is said to be *blocked* if it is unable to communicate with the coordinator or some other non-uncertain participant. A blocked participant can neither commit nor abort without violating the ACP, and will therefore have to wait indefinitely (until it receives instructions on whether to

[1] *Atomicity and durability are automatically enforced, but as a direct consequence of non-serialisability special precautions may be required to ensure consistency.*

The two phase commit protocol

Whenever a transaction spans two or more nodes in a distributed system, one must ensure that either all involved nodes commit the work of that transaction or all involved nodes abort the work of that transaction. This is known as atomic commitment and is ensured by means of so-called atomic commit protocols (ACPs). The ACP which is almost always used in practice is two phase commit (2PC), the basics of which is outlined below.

One of the nodes participating in the execution of a distributed transaction must be designated as the coordinator. This could e.g. be the node where the transaction execution was initiated. All other nodes are referred to as participants (a node can be the coordinator for one transaction and a participant in another). Unless something out of the ordinary has happened, forcing the transaction to abort, the coordinator will at some point in time start making preparations for commitment of the transaction's work. It will then enter phase one of the 2PC protocol. This is illustrated in Figure 1.

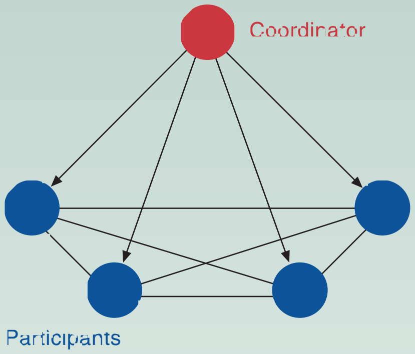


Figure 1 In the beginning of phase one the coordinator sends a vote request message to each participant. The semantics of this message can be paraphrased as "I'm willing to commit, how about you?"

A participant that receives a vote request message and is ready to cast its vote, must vote either Yes or No. This takes place in the second part of phase one and is illustrated in Figure 2.

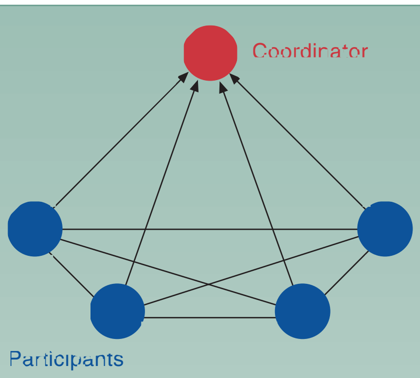


Figure 2 At the end of phase one the coordinator receives a Yes or No vote from each participant. The moment a participant votes Yes, it gives up its right to unilaterally abort; from then on it must wait until it receives a final Abort or Commit message. Such a participant is referred to as "uncertain", and some failure scenarios can cause it to remain blocked indefinitely

If all participants vote Yes, the coordinator will decide to commit the transaction. If at least one participant votes No, the coordinator has no other option than to abort the transaction. As soon as all votes have been collected, the coordinator enters the second phase of the 2PC protocol and informs the participants of the final outcome. This is illustrated in Figure 3.

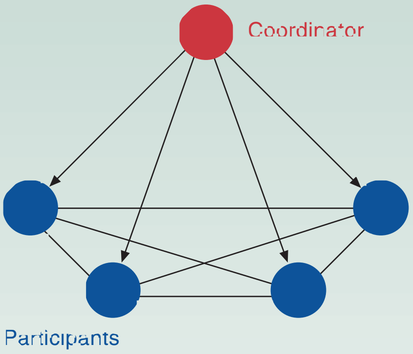


Figure 3 In the second phase the coordinator sends an Abort or a Commit message to each participant. If they all vote Yes it will be a Commit message, otherwise an Abort message

commit or abort). It has been proved that blocking free ACPs cannot be found [21]. 3PC is less prone to blocking than 2PC but is seldom used [4].

Thus, it is interesting to investigate ways of minimising the problems caused by blocking. Two main approaches are possible: (1) Violating the ACP by allowing blocked participants to unilaterally make an abort/commit decision. (2) Retain transactional control over all accessed resources until a proper abort/commit decision can be reached. Both approaches cause problems, and a choice between them can only be made by weighing their pros and cons.

The first approach is taken by [15]. This inevitably makes it necessary to rerun incorrectly aborted transactions and to compensate for incorrectly committed transactions, none of which is trivial in general. Moreover, such approaches will necessarily cause the database to be inconsistent for some time after each discrepant abort/commit decision. This could potentially cause severe problems for other transactions, unless they can distinguish between ordinary (i.e. reliable) data items and those that are the results of tentative commit decisions. Thus, it is briefly mentioned that approaches of this category would benefit by having data values left behind by prematurely committed transactions be annotated (before committing) with Apotram style reliability indicators. This would alleviate the problems caused

(Continued from previous page)

As soon as the messages are received by the participants they do as they are instructed and then terminate. The coordinator terminates after sending out all messages. Thus, the 2PC protocol ensures that either the coordinator and all participants commit or that they all abort, i.e. it ensures atomicity of the abort/commit decision. That's why it's called an ACP.

There are many things that can go wrong during the life of a distributed transaction, both before and during execution of the 2PC protocol. Such problems can be divided into two basic categories, node and communications failures. This is illustrated by Figure 4 and 5.

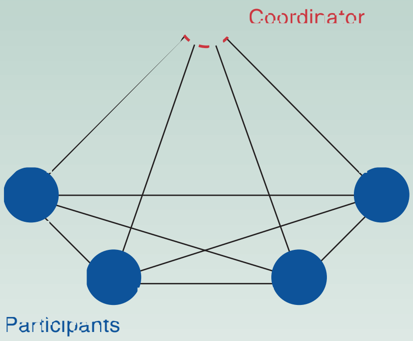


Figure 4 A node failure means that one of the nodes is not working properly. In particular it cannot participate in the 2PC protocol

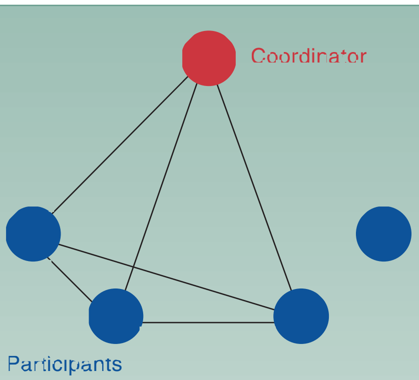


Figure 5 A communications failure can cause the network to become partitioned, i.e. divided into two or more segments that are unable to communicate with each other

If e.g. the coordinator fails just after all participants have voted Yes in the first phase of 2PC but before the final Commit message has been sent out, i.e. the coordinator fails when the participants are uncertain, then all the participants will be blocked until the coordinator starts functioning properly again. And if a participant is cut off from the other nodes by a communications failure just after it voted Yes in the first phase of 2PC, it will be blocked until communications are resumed and it learns about the final outcome of the transaction.

by having an inconsistent database since other transactions would be informed of the quality of data they happen to encounter, but it would not help at all with the inherent problems of compensation.

This paper advocates the second approach. For the sake of the discussion it is assumed that the concurrency control mechanism is locking and that a blocked transaction will keep its locks until a proper commit/abort decision is reached. Thus, atomicity is not compromised, a transaction's work can be rolled back at any time before commit/abort (although uncertain transactions cannot make a unilateral decision), and there is no need for compensating (trans)actions. The challenge then is to make the indefinite holding of locks as small a problem as possible. The Apotram solution is to use parameterized access modes.

A blocked Apotram transaction can modify its own concurrency level by choosing more lenient share and/or protection levels, the details of which would be application dependent. Two kinds of uncertainty of the data locked by such a transaction are of particular importance in this context; the probability that the transaction will eventually be committed, and the probability that further modifications of locked data items will take place once communication with other nodes is re-established.

As an example, let us say that an analysis of possible failure scenarios shows that blocked transactions will belong to one of the following categories:

1 At least 80 % chance of eventually committing, and data values are unlikely to change.

2 At least 80 % chance of eventually committing, but data values are likely to change.

3 Between 50 % and 80 % chance of eventually committing, and data values may be likely to change.

4 Less than 50 % chance of eventually committing.

These situations represent four different classes of reliability for uncommitted data. It therefore makes sense to define one reliability indicator for each of them, let us call these u1, u2, u3, and u4, respectively. It is assumed that each database in the DDBS has a metadatabase containing, among other things, a set of defined reliability indicators I (along with a description of their semantics) of which u1 to u4 would be a subset. Further, it makes sense to let the domain of access mode parameters D be a subset of I (I will in general also contain reliability indicators that are unrelated to transactions; such reliability indicators do not belong in D). Let us assume here that u1, u2, u3, u4 ∈ D. Thus, a transaction that discovers that it is blocked, could modify its write mode and thereby cause its write locks to be downgraded from W to W(u1), W(u2), W(u3), or W(u4), depending on which of the above identified reliability categories the application at hand belongs to. Note that the latter is predefined, and need not be determined at run time.

The only way for a transaction to discover that it is blocked, is to try to communicate with other nodes and then fail to receive any messages from them. Such a discovery must be based on some predefined time-out limit, which must be chosen based on heuristics. It is therefore worth pointing out that a transaction need not wait until it discovers that it has been blocked before it can choose a more lenient share level. Rather, transactions that are part of distributed transactions can make sure the desired downgrading of locks takes place as soon as they complete their work. That way, they will cause less problems for other transactions while they wait for the vote request and Commit/Abort messages from the co-ordinator, and they will be prepared in advance for a possible blocking situation.

3.2 Waiting versus aborting

The entire discussion of recovery and atomic commit protocols for DDBSs in [4] seems to be based on the tacit assumption that hardly anything is more important than to minimise the time resources are locked. As a consequence, transactions are sometimes aborted even though this is not necessary. In other words, the attitude seems to be that a transaction should be aborted if there is reason to believe that it *may* not be possible to commit it. An example can be found in the discussion of the Cooperative Termination Protocol for 2PC (page 228): if *q* receives a decision request from *p*, and *q* has voted No or not yet voted Yes, then *q* sends an Abort message back to *p*. This implies that a participating transaction *q* which is active and unaware of any problems, can suddenly reach a unilateral abort decision (affecting the entire distributed transaction) just because some other participating transaction *p* had timed out waiting for a message from the co-ordinator. This argument is further developed below.

This may be a good strategy with traditional transactions, but if holding locks for an extended period of time was not such a big problem, other solutions might be worth considering. For example, the termination protocol could be redesigned to allow time for replies from all participants to be sent back to the initiator, rather than letting any responder who is able to cause an abort do so. That way, a decision could be based on a more complete picture of the situation, instead of the pessimistic approach just outlined. Generally speaking, since parameterized read and write modes do not reduce concurrency the way classical write and read modes do, their use in distributed transactions would make more elaborate commit, termination, and recovery protocols viable. Such protocols would require more elapsed time, but would cause fewer aborts.

One observation is of particular interest in this context. Consider a distributed transaction with a coordinator and two or more participants. Let us say that the coordinator sends out vote requests and all participants but one reply Yes, and that the participant that does not reply Yes just has not completed its work (so it has not cast its vote yet). If this situation persists for some time, either the co-ordinator will time out or one of the participants waiting for the final Commit/Abort message from the co-ordinator will time out and send a decision request to all other participants. Given a pessimistic termination protocol (i.e. with a bias towards aborts), the transaction manager at the non-voting node would, unless special precautions were taken, decide to abort if it received a decision request from another participant. Thus, even though there are no node or communication failures, the transaction is aborted. This is an example of a situation where a more optimistic, although more elaborate, termination protocol could be an advantage.

A possible objection to this line of reasoning is that there must be some limit to how long participants should wait for each other, and that this is merely a question of choosing appropriate time-out limits. However, such an argument misses the point that the node at which the elapsed time measurement is performed has little or no knowledge of conditions at the other nodes. Problems that have to do with one participant getting into livelock, deadlock, or unusual waiting for local resources can be better handled locally. In other words, one participant cannot in general know whether another participant has become stuck, in which case it should be aborted, or just happens to be tardy, in which case it might be better to let it run to completion.

Considering once more the CTP for 2PC of [4, page 228], one wonders why no attempt is made to distinguish between tardy and stuck participants. The only reasonable answer seems to be that with resources locked in ordinary read and write modes at the other participants, one cannot afford to spend more time waiting whether the problem is one or the other. Since Apotram enables transactions to wait with a lower concurrency penalty, it has advantages to offer DDBSs; the use of more elaborate (and therefore time consuming) termination protocols becomes feasible, and one can afford to use more generous time-out limits, thus making it possible to reduce the number of aborts.

3.3 Three phase commit

3PC is less prone to blocking than 2PC but is seldom used in practice. The reason is that blocking is not a serious enough problem to justify the extra costs associated with 3PC; more message passing and later freeing of resources. The increased message traffic can be compensated by installing more communication bandwidth, so the main problem with 3PC is that it takes more time than 2PC. Since, as argued above, waiting is less of a problem with Apotram this means that 3PC becomes a more viable option. This is particularly true for applications that do not involve end-user interaction.

3.4 Waiting in general

It is an intrinsic property of atomic commit protocols in DDBSs that they lead to waiting. The only way a node A can discover that it is unable to communicate with node B (whether B is down or communication between A and B is broken) is by using a time-out mechanism, which requires waiting. Even if there are no node or communication failures, execution at the nodes involved in a distributed transaction will almost certainly complete at different times. Hoping to cause subtransactions to complete at the same time, one could try to start them at different times according to their expected duration. But in general, it is very difficult to predict elapsed times for subtransaction executions, since they will be influenced by dynamically varying workloads on communication links and nodes. Thus, all participant transactions except one have to passively wait for some time before their work can be committed or rolled back. For distributed transactions that allow general nesting there will be even more waiting involved, since a parent in general cannot respond to a prepare message without first soliciting its children [17]. Provided parameterized write modes are used as discussed above, this unavoidable waiting is less of a problem with Apotram than with the classical transaction model.

3.5 Application-oriented transaction management

A central idea in Apotram is that transaction management should be tailored to the specific needs of different applications [1], a concept that applies in this context, too. Consider e.g. the standard notion that a participant which fails before casting its vote should unilaterally abort when it recovers. This is motivated by the desirability of being able to perform independent recovery, but is strictly speaking not necessary. Assuming that the participant in question had completed and logged its work before failing (or was at least able to continue where it left off), it would have been possible to let the co-ordinator and the other participants wait for recovery to take place and then complete the commit protocol together. Given that recovery can take a long time, this way of doing things is surely unacceptable for many application domains. However, the key question is whether there are application domains for which such a protocol would be attractive. Good candidates can be found in telecommunications were e.g. routing tables need to be updated at preset points in time. Such tables must be continuously available for traffic routing but at the same time it would be desirable to perform the updates with proper transactional control [14][2].

Generalising the argument, several situations can arise where standard commit, termination, or recovery protocols [4], [18] choose to abort even though one is not forced to do so. In other words, abort is chosen because it is *reasonable* rather than *necessary.* The latter is an important observation for anyone who needs to understand the *semantics* of distributed transactions (note that Presumed Abort and Presumed Commit protocols [17] are optimisations of 2PC, and thus without semantic significance). It is well known that application domains exist for which classical transactions are unsuitable [12], [6], [13], and the semantics provided by the just mentioned protocols simply represent a special case. Thus, transaction models that make it reasonable to relax some of the (tacit) assumptions that underlie various protocols found in current DDBSs, could be a key to make true DDBS functionality a viable alternative to the ad hoc solutions used in several kinds of distributed computing today, e.g. in telecommunications.

3.6 Reliability of uncommitted data

Access to uncommitted data is in general less attractive the less reliable the uncommitted data is. The single factor that contributes the most to making uncommitted data unreliable, is probably that it is subject to rollback. It is therefore interesting to observe that the frequency of aborts can be reduced if Apotram is being used. As discussed above, Apotram makes waiting a feasible alternative to an abort. In the extreme case, one could choose to set all time-out limits to infinity, allowing transactions to wait indefinitely and thus reducing significantly the probability of aborts. The point being made is that Apotram cannot only allow access to uncommitted data in a controlled manner, but can also make uncommitted data more reliable by reducing the rollback frequency, thus making access to uncommitted data more attractive.

3.7 A multidatabase remark

A basic problem with multidatabases is that conflicting interests exist between owners of multidatabase nodes and users of multidatabase transactions [7], [15], [18], [5]. In particular, any transaction that spans multiple nodes could in principle benefit from being allowed to execute an atomic commit protocol (ACP), e.g. two-phase commit, but this cannot always be reconciled with the autonomy requirement of multidatabase nodes. [19] argues that 2PC participation should not be seen as a violation of node autonomy. Nevertheless, the picture is not black and white here but contains shades of grey; the degree of autonomy required by participating nodes may vary from one multidatabase system to another. It is therefore interesting to observe that Apotram makes it possible to minimise the impact of holding locks, and this means that it might be acceptable in some multidatabase systems to allow transactions to execute ACPs, provided use of parameterized lock modes can be enforced. It is also worth noting that the availability of parameterized lock modes can make new kinds of multidatabase applications acceptable. A further discussion of this topic is beyond the scope of this paper.

4 Apotram and distributed applications

So far in this paper, we have been concerned with how protocols for commitment and recovery can be written in order to exploit certain features of Apotram. This section will explain how the same features can also be exploited by applications that are designed to run on a distributed database system.

4.1 Description of problem

Consider a distributed database and a transaction that needs to access data on multiple sites. Whether or not data replication is in use, is without interest to the problem discussed here. Assume that this transaction is to be run at some predefined point in time, or just that it is to be run as soon as possible once it is initiated. Such transactions can be found in several application domains; this paper will use a distributed version of the traditional suppliers and parts database for illustrative purposes. If at least one of the nodes that need to be updated is unavailable (due to node or communication failure), classical transactions offer, as mentioned in the introduction, no options other than aborting or waiting, both of which are unacceptable. Aborting is unacceptable because the nodes that are available must be updated immediately in order to ensure maximum service availability. Waiting is unacceptable for the same reason, i.e. the updated resources at the available nodes must be committed such that they can be accessed by other transactions.

Faced with two unacceptable options, designers of distributed applications are forced to do something else, i.e. they have to resort to some kind of ad hoc solution. This typically means performing the necessary operations with little or no transactional support. One possible solution could be to use one independent transaction per node that is to be updated. This would make it

[2] *I am referring here to what H.V. Jagadish of AT&T Bell Labs said in his SIGMOD tutorial, not to what can be found in the one page summary in the proceedings.*

the responsibility of the application to ensure that all necessary transactions are executed and committed, and should there be any dependencies between operations on different nodes, the application would have to handle that too. Another possible solution could be to use open nested transactions [13]. This would alleviate the application designers somewhat compared to the independent transactions case. A third possible solution, which seems to be more common than one would like to think, is to forget about transactions altogether.

The three just sketched solutions have two things in common: they all make it possible to make data available "right away", and they all sacrifice atomicity and thereby recoverability in order to do so. In other words, there is a need to compromise isolation but because traditional transactions cannot selectively use the ACID properties, all four of them are sacrificed. Thus, if the need to roll back committed or non-transactional work arises, this must be explicitly handled in the applications by means of compensating (trans)actions.

4.2 Distributed Apotram transactions

Assuming that all nodes in the DDBS have support for Apotram, how could transactions that update multiple nodes be handled in the presence of node or communication failures? The answer is that the solution could be tailored to the application at hand. This should be done by defining reliability indicators and access modes that correspond to situations that may arise. Let us say that an analysis of possible failure scenarios results in a need to distinguish between the following cases:

1. Data at a subset of the nodes has been updated and will be committed a soon as the remaining nodes become available, except for the unlikely event that a consistency constraint violation is detected, in which case the transaction will abort.
2. Same as 1 but with the added complication that already updated values may be further modified based on information from the not yet accessed nodes.
3. Data at a subset of the nodes has been updated but has a significant risk of rollback and must therefore be regarded as unreliable.

These situations represent three different levels of reliability for uncommitted data. It therefore makes sense to define one reliability indicator for each of them, let us call these v1, v2, and v3, respectively. Again, it is assumed that each database in the DDBS has a metadatabase containing, among other things, a set of defined reliability indicators (along with a description of their semantics) of which v1, v2, and v3 would be a subset. As above, reliability indicators may be used as access mode parameters.

In order to make the example more concrete, let us assume that the DDBS in question is run by some chain X of supermarkets operating in several countries within the European Economic Community (EEC). X has multiple warehouses and supermarkets in each country where it is present, and does business with a large number of suppliers all over the EEC. Each supermarket and warehouse has its own database which acts as a node in an EEC-wide DDBS. In addition to that, each national headquarters as well as the European headquarters has a database that is also a part of the DDBS. Certain goods are always ordered by the supermarkets directly from the suppliers, e.g. milk and bread. Other goods are usually delivered to warehouses, from where they are ordered by supermarkets as needed. However, such goods may be ordered by supermarkets directly from suppliers in urgent situations. All business transactions of X involving supermarkets, warehouses, and suppliers are subject to certain corporate rules and regulations, the purpose of which is to prevent uncontrolled spending of money, excessive storage of goods at warehouses, cannibalistic competition among supermarkets, etc. The examples are meant to be illustrative rather than realistic.

4.2.1 Case 1

Say a warehouse is getting low on some kind of goods. A deal is negotiated with a supplier, and an entry in the warehouse database is made indicating date of delivery and amount ordered. Unfortunately, the national headquarters' database is currently down, and the transaction cannot be completed until certain controls against this database have been performed. However, this transaction is such that only extreme events at the national or European level could cause it to be incompatible with corporate policy. Thus, the delivery in question will almost certainly take place, and information about it is useful for supermarkets and warehouses that continue to access the database. Instead of aborting the transaction and trying again later, the share level of the updating subtransaction can be changed such that the locks are downgraded from mode W to W(v1). This will cause a reliability indicator of type v1 to be attached to the data items in question, and readers accessing the database in a suitable parameterized read mode, will know that the information is quite reliable.

4.2.2 Case 2

Say the national management of X in a country becomes aware that a competing chain of shops will start a campaign next week with major discounts on product P1. There is an urgent need to react to this situation, and the various supermarket managers must be given information about how to prepare as soon as possible. National management decides to launch a campaign for product P2, and each supermarket will receive large deliveries of P2 in the near future.

Unfortunately, communications has been broken with the western part of the country where a large warehouse and two supermarkets are located. In order not to lose time, a distributed transaction updating the databases of all available supermarkets and warehouses is started. Through this transaction, warehouses are told how much of P2 they must be ready to ship when, and supermarkets are told how much of P1 they can hope to receive when. There is hardly any reason to doubt that this distributed transaction will eventually be committed some time after communication with the western regions are re-established. However, once that happens, adjustments of amounts and times of deliveries may be necessary. Thus, locks on updated items can be downgraded from W to W(v2), with accompanying automatic insertion of v2 reliability indicators. In case of data value adjustments, W(v2) locks should temporarily be upgraded to W in order not to cause problems for parameterized readers.

4.2.3 Case 3

Say that a certain product is facing severe competition from a new product that has been introduced on the market in one country. National management decides to do a wholesale operation with the complete stock of the old product and purchase enough of the new product to replace the old. Given that the market value of their old stock is low, this will require a significant investment which must be checked with the European headquarters, the database of which is currently unavailable. In order to prevent warehouses from shipping anything of the old product, a distributed transaction is run against all warehouses marking the entire stock as sold. Since there is a significant risk that the whole plan may have to be reworked, however, the updated items in the warehouse databases are locked in W(v3) mode (and thus annotated with v3 reliability indicators) while one is waiting for contact with the European headquarters to be re-established.

5 Conclusions

Communication and node failures are inescapable ingredients of DDBS operations, and can cause blocking as well as transaction aborts. 3PC is less prone to blocking than 2PC, but is not used in practice. For applications that can accept the increased elapsed time incurred with 3PC, this ACP could be made more acceptable with Apotram. However, my primary approach to blocking is not to make it less likely, but to make it less of a problem. This is achieved by providing customisable reliability indicators and access modes that can be tailored to meet the needs of different application domains, thus making it possible for a blocked transaction to share its locked data with others. In addition to making blocking a less severe problem, Apotram makes more elaborate commit, termination, and recovery protocols, as well as more generous time-out limits, feasible, which means that transaction aborts can be avoided in some situations.

Reduced abort frequency has performance and resource consumption benefits. Firstly, waiting transactions can be run to completion as soon as the problem in question is removed. Thus, it may be completed at the earliest possible time. Secondly, if a transaction is forced to abort rather than allowed to wait, it will or may have to be restarted later – possibly multiple times – which would represent a waste of computing resources.

Waiting is an intrinsic part of distributed transaction management, and it therefore makes sense to investigate transaction models where waiting transactions have minimum impact on their surroundings. This paper has shown that Apotram can be used for such purposes.

Further work is necessary (1) to identify application domains for which the proposed semantics are suitable, (2) to design commit, termination, and recovery protocols that exploit Apotram features, and (3) to develop heuristics that are necessary for proper use of such protocols.

References

1 Anfindsen, O J, Hornick, M. *Application-oriented transaction management.* Kjeller, Norwegian Telecom Research, 1994. (Technical report TF R 24/94.) (http://www.nta.no/brukere/olea/pub/).

2 Anfindsen, O J. Dynamic cooperation between database transactions by means of generalized isolation levels. In: *Proceedings of 2nd international conference on concurrent engineering : research and applications.* Virginia, McLean, 1995, 249–260.

3 Anfindsen, O J. *Apotram : an application-oriented transaction model* (preliminary title). PhD thesis, UNIK center for technology at Kjeller, Department of Informatics, University of Oslo, 1996. Forthcoming.

4 Bernstein, P A, Hadzilacos, V, Goodman, N. *Concurrency control and recovery in database systems.* Reading, Mass., Addison-Wesley, 1987.

5 Breitbart, Y, Garcia-Molina, H, Silberschatz, A. Transaction management in multidatabase systems. I : *Modern database systems,* Kim, W (ed). Reading, Mass, Addison-Wesley, 1995, 573–591.

6 Elmagarmid, A K (ed.). *Database transaction models for advanced applications.* San Mateo, Calif., Morgan Kaufmann Publishers, 1992.

7 Elmagarmid, A K, Pu, C (eds). Special issue on heterogeneous databases. *ACM Computing Surveys,* 22, (3), 1990.

8 Farrag, A A, Özsu, M T. Using semantic knowledge of transactions to increase concurrency. *ACM transactions on database systems,* 14, 503–525, 1989.

9 Garcia-Molina, H. Using semantic knowledge for transaction processing in a distributed database. *ACM transactions on database systems,* 8, (2), 186–213, 1983.

10 Georgakopoulos, D, Hornick, M, Krychniak, P. An environment for the specification and management of extended transactions in DOMS. In: *Proceedings of the third international workshop on RIDE-IMS'93,* Vienna, Austria, 1993, 253–257.

11 Georgakopoulos, D et al. Specification and management of extended transactions in a programmable transaction environment. In: *Proceedings of the 10th international conference on data engineering,* Houston, Texas, 1994, 462–473.

12 Gray, J. The transaction concept : virtues and limitations. In: *Proceedings of the 7th VLDB conference,* 144–154, Cannes, France, 1981, 144–154.

13 Gray, J, Reuter, A. *Transaction processing : concepts and techniques.* San Mateo, Calif., Morgan Kaufmann Publishers, 1993.

14 Jagadish, H V. Databases for networks : tutorial summary. In: *Proceedings of ACM SIGMOD international conference on management of data,* Minneapolis, Minn., 1994, 522.

15 Levy, E, Korth, H F, Silberschatz, A. A theory of relaxed atomicity. In: *Proceedings of 10th ACM symposium on principles of distributed computing,* 1991, 95–109.

16 Lynch, N. Multilevel atomicity : a new correctness criterion for database concurrency control. *ACM transactions on database systems,* 8, 484–502, 1983.

17 Mohan, C, Lindsay, B. Efficient commit protocols for the tree of processes model of distributed transactions. In: *Proceedings of the 2nd ACM symposium on principles of distributed computing,* 1983, 40–52.

18 Özsu, M T, Valduriez, P. *Principles of distributed database systems.* Englewood Cliffs, New Jersey, Prentice-Hall, 1991.

19 Pu, C, Leff, A, Chen, S-W F. Heterogeneous and autonomous transaction processing. *IEEE Computer,* 24, (12), 64–72, 1991.

20 Shillington, J, Özsu, M T. Semipermeable transactions and semantics-based concurrency control for multidatabases. In: *Proceedings of the third international workshop on RIDE-IMS'93,* Vienna, Austria, 1993, 245–248.

21 Skeen, D, Stonebraker, M. A formal model of crash recovery in a distributed system. *IEEE transactions on software engineering,* 9, (3), 219–228, 1983.

22 Soparkar, N, Silberschatz, A. Data-value partitioning and virtual messages. In: *Proceedings of the 9th ACM SIGACT-SIGMOD-SIGART symposium on principles of database systems,* 1990, 357–364.

Status

International research and standardization activities in telecommunication

Editor: Endre Skolt

Introduction

BY ENDRE SKOLT

EURESCOM[1] is one of the most important instruments for telecommunications research in Europe, and has shown to be very successful for Telenor. In 1995 Telenor R&D participated in 19 projects with a total budget of 180 man months. For 1995 EURESCOM's total project budget was 1275 man months, having participants from 24 shareholders. In this issue of *Telektronikk*'s Status section we will look closer at 4 projects that are linked to the study areas Teletraffic and dimensioning, IN, and TMN.

In the first paper ,"Migration from today's IN, TMN and B-ISDN towards TINA[2]", Mr Rødseth presents ongoing activities in project 508. TINA, which is proposed to be a future platform for provisioning, management and execution of telecommunications services may in a few years replace the current IN and TMN standards. The focus of P508 is the challenge to migrate existing IN- and TMN-standards towards TINA.

The next paper by Mr Jensen, reviews the status of the project "Methods and Specifications for Tools to dimension Intelligent Networks". As the number of services offered by the IN-platform are increasing, and consequently the traffic load, there is a need for appropriate tools to dimension network elements and communication links. Special items that are addressed in this paper are dimensioning processes, load and mechanisms, and technical planning of IN deployment.

The third paper is on project 507 "Mobility Applications Integration in IN". This project covers activities such as support of mobile services on future IN architectures, evolutionary paths from existing telecommunications systems towards the 3rd generation mobile systems, and development of a suitable information model for relevant mobile applications and services. The author is Mr Lauritzen.

EURESCOM has, since it was established, launched several projects in the TMN area addressing the problem of management of pan-European networks and services. Up till now these projects have produced theoretical results such as analyses, guidelines, models and specifications. In order to increase their value, project 408 was launched to do testing, validation and experimentation of existing results. Mr Kristiansen presents a comprehensive outline of the project and preliminary results.

Table 1 List of previous contributions to the Status section

Issue No.	Title of published papers	Authors
4.93	Service definitions	Ingvill H. Foss
4.93	Radio communications	Ole Dag Svebak
4.93	Transmission and switching	Bernt Haram
4.93	Intelligent Networks	Endre Skolt
4.93	Broadband	Inge Svinnset
1.94	Terminal equipment and user aspects	Trond Ulseth
1.94	Signal processing	Gisle Bjøntegård
1.94	Telecommunications Management Network	Ståle Wolland
1.94	Teletraffic and dimensioning	Harald Pettersen
1.94	Data networks and open system communications	Berit Svendsen, Mette Røhne
2.94	The TINA Consortium	Tom Handegård
2.94	Telecommunications languages and methods	Arve Meisingset
2.94	Message Handling Systems	Geir Thorud
2.94	Security	Sverre Walseth
3.94	EU's research programme ACTS	Eliot J Jensen
3.94	UPT- Service concept, standardisation and the Norwegian pilot service	Frank Bruarøy, Kjell Randsted
3.94	Status report form ITU-TSB, SG 1	Elisabeth Tangvik
3.94	Future mobile communications	Ole D Svebak
3.94	The CIO project	Gabriela Grolms
4.94	Eurescom work on ATM network studies and the European ATM pilot network	Inge Svinnset
4.94	Terminal equipment and user aspects	Trond Ulseth
4.94	Telecommunications Management Network	Ståle Wolland
2/3.95	Documents types that are prepared by ETSI	Trond Ulseth
2/3.95	ATM traffic activities in some RACE projects	Harald Pettersen
2/3.95	The Public Network Operator Cipher project	Øyvind Eilertsen

[1] *EURESCOM: European Institute for Research and Strategic Studies in Telecommunications. (For more information on EURESCOM, see Telektronikk 2.94).*

[2] *TINA: Telecommunication Information Networking Architecture.*

Migration from today's IN, TMN and B-ISDN towards TINA

BY THOMAS RØDSETH

This article presents the work done in EURESCOM project P508 in 1995 where Telenor R&D has been participating. The result of the work described are available on request to EURESCOM.

Background

Major research efforts are currently invested by several telecommunication manufacturers, operators, and large players in the computer industry, to define and validate a software architecture that will enable the efficient introduction, execution and management of new and sophisticated telecommunication services. This initiative is called TINA – Telecommunication Information Networking Architecture – and the work is driven by a core team called TINA-C which consists of 40 experts from the TINA member companies. In addition, several auxiliary projects in home companies are supporting the core team by using and evaluating their work. The core team started its work in 1993 and will finish its specifications of the TINA architecture by the end of 1997. The specifications issued by TINA-C are starting to become more stable, but still several issues remain uncovered [2] [3] [4].

TINA is seen as a future architecture that can replace the current IN [10] and TMN [5] standards, and create a common framework and platform for what is today two different standards. The potential seems to be most evident for the coming broadband networks, for which IN standards are still not defined. TINA-C is defining an idealistic architecture, in the sense that it is not restricted to conform with current legacy systems, although several of the current telecommunication and computer standards (like TMN [5], GDMO [6], ODP [7]) are taken as input and components of the TINA specifications. This leaves a gap between current standards, legacy systems, and TINA. The ITU-T and ATM-Forum standards for broadband signalling are issued, and now penetrating the market for broadband switching equipment. These standards are addressing several of the aspects also addressed by TINA, and how they should interact – or be replaced – are still not studied internationally.

Table 1 Accrual of business benefits

Business Benefits	Customers	Players*	Industry	National Economy	Regional Economy
Reduced Costs	2	1	3		
Revenue Opportunities		1	2	3	
Time to Market	2	1	3		
Creating Efficiencies	2	1			3
Increased Competition	1		2	3	
Enhanced Customer Benefits	1			2	3
Enhanced European Economy			3	2	1

* *The group considers the players as Service Provider, Network Provider and Content Provider*

The Eurescom project 508 – Evolution, migration and interworking towards TINA – focuses on these challenges, proposing how to migrate towards, and interact with, TINA from current standardised architectures like IN, B-ISDN and TMN. The work is done in collaboration with TINA-C, giving corrections and input to the core team. Most results of P508 will be publicly available and possibly presented to standardisation bodies.

P508 was initiated by CSELT (Italy), and started at the beginning of 1995. The project period is two years. The project has divided the work for 1995 into 4 workgroups:

- **Business Aspects** – which looks at the business benefits of introducing TINA in different European markets
- **Network Aspects** – which proposes network scenarios for migration from today's narrowband networks and IN, and from the recently standardised B-ISDN networks
- **Service and Network Management Aspects** – which analyses available architectures for management of networks and services, and studies migration from these architectures towards TINA
- **Information services** – which looks at information services, and how important features of these services on the Internet could be offered by TINA.

The results from the four workflows will be summarised and consolidated into a final harmonised paper [1] that covers all the above aspects of migrating towards TINA.

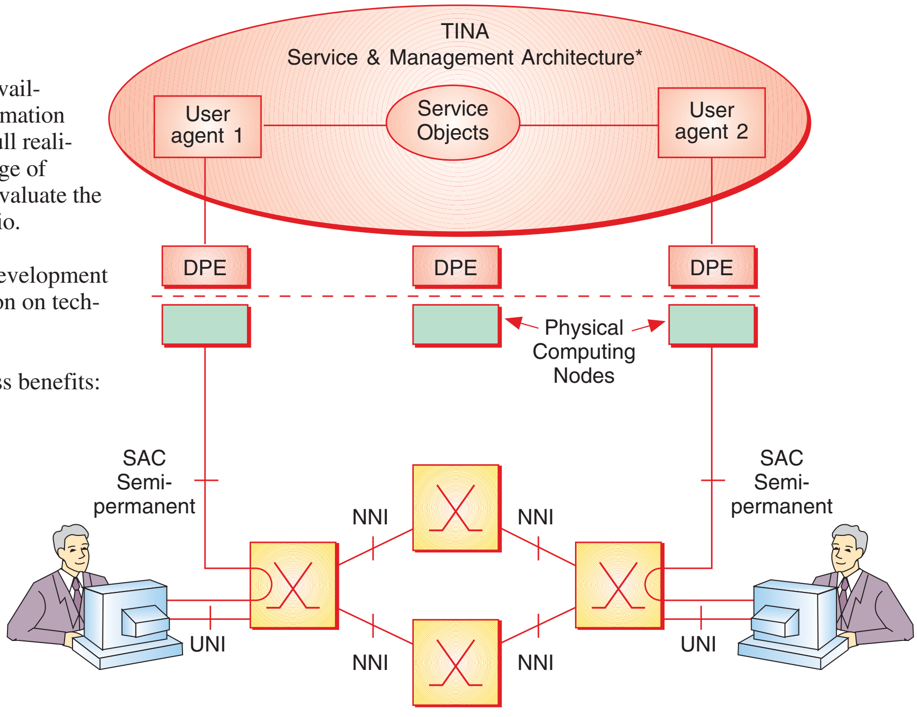
Work carried out in 1995

The first official deliverable from the project is due at the end of this year, and the work carried out so far by the different workgroups are summarised below. Telenor is participating in the group for Network Aspects, which consequently will be given most attention in the description. Some application cases were selected for each workgroup, to evaluate the migration and interworking scenarios identified. These evaluations are still not complete, and consequently not covered below.

Business benefits

The group has adopted the following methodology to evaluate the business benefits provided by TINA under specific market conditions. The initial step was to identify and define the technical capabilities that a TINA implementation will deliver. Separately, but not in parallel, a set of generic business benefits were generated. The output from these stages allowed an assessment of how the capabilities may give rise to business benefits. Each of the technical application cases were then developed and analysed to determine how a particular scenario may further accentuate a potential business benefit.

As the technical scenarios may take place in a number of different environmental conditions, an investigation must be under-taken of both regional market conditions and of the available technology base. These stages will provide information with which to assess what constraints will limit the full realisation of the potential business benefits. The final stage of the process, having completed this assessment, will evaluate the opportunities and threats associated with each scenario.



* The TINA service & management architecture is physically distributed over several computing nodes

Figure 1 Scenario BB-1

This process can be iterative. For example, further development of technical scenarios could provide more information on technical capabilities.

The group so far has identified the following business benefits:

1 Reduced cost

2 Revenue Opportunities

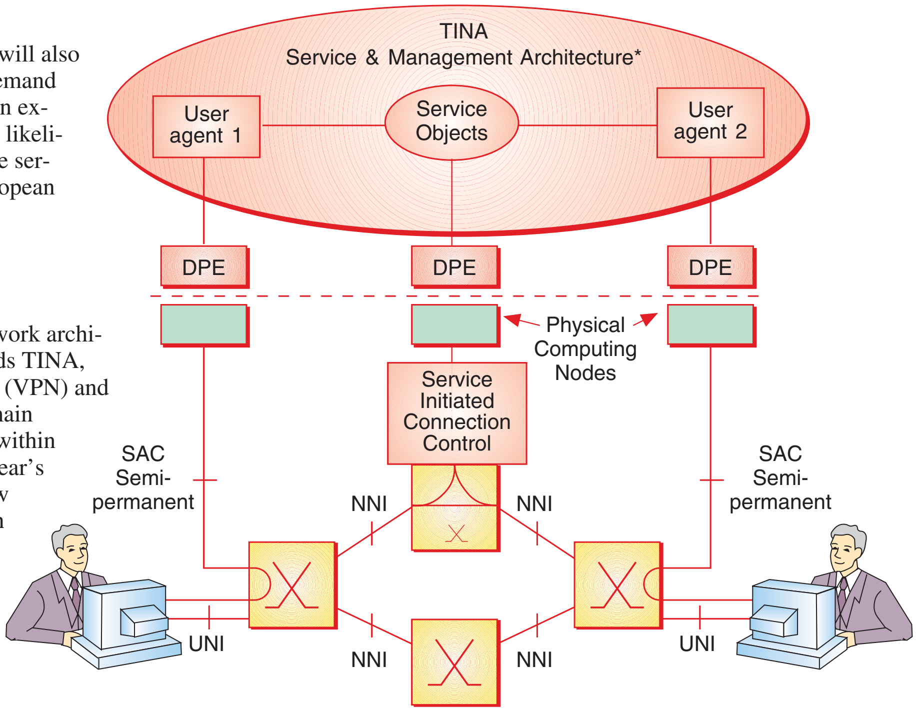
3 Time to Market

4 Increased Competition

5 Customer benefits.

The benefits were then evaluated and graded for the different parties. The results are summarised in Table 1, illustrating how the business benefits are likely to be weighted across the different levels. These are ranked as first, second, or third in terms of likely benefit accrual. Those with no marking fall outside this scale. This shows, for example, that cost reduction is likely to benefit the players most, whilst increased competition is likely to benefit the customer most.

In order to examine the impact of TINA, the group will also investigate the three application cases Movie-on-Demand (MoD), Broadband VPN, and Telebanking (a chosen example of information services). This to estimate the likelihood of their success and how TINA can serve these services better than other technologies in different European markets.



* The Tina service & management architecture is physically distributed over several computing nodes

Figure 2 Service Initiated Connection Control scenario

Network aspects

This group has studied migration from existing network architectures both narrowband/IN and broadband, towards TINA, using the application cases Virtual Private Network (VPN) and Multi-media Multiparty Conference (MMC). The main focus in 1995 has been to address how an operator within his own domain can migrate towards TINA. Next year's work will use these migration scenarios and see how they interwork with other non-TINA networks, both broadband and narrowband/IN.

The work on broadband networks has so far identified two migration scenarios with B-ISDN [8] CS-1 as the starting point. The scenarios are based on having established the TINA service and management architecture "on top of" the standardised B-ISDN network. In both scenarios the users are connected via semi-permanent links to the service and management objects in TINA, which make up the service objects (see Figure 1). We have called this connection the *Service Access Connection* (SAC). It allows the users to send their service requests directly to the service objects, and to freely interact with the service objects throughout the service session. Both scenarios avoid the IN concept of trigger points and Basic Call State Model by allowing the user direct access to the service objects.

The first broadband scenario – BB-1 – suggests a way to introduce TINA as the service architecture without changing any of the existing B-ISDN standards. A service request from the user is sent to the TINA service objects through the SAC. The request is then processed, and the end-to-end parameters are negotiated with all end-user terminals through their SACs. When a connection is needed, the service objects contact one of the participating terminals via their SAC, to further request the connection set-up. The terminal now establishes the needed connection(s) by translating the request to B-ISDN signalling messages, sent over their User to Network Interface (UNI).

In scenario BB-2 the user accesses the service objects in the same manner as in BB-1. However, scenario BB-2 requires changes in B-ISDN signalling standards, to provide the TINA service and management objects more control of the connection set-up. The necessary connections are now requested directly from the switching equipment, and not through any of the users. To be able to do so, we suggest to have an interface in one or more switches that are able to set up connections using B-ISDN signalling towards all participating users and necessary network resources (see Figure 2). This interface we have called the *Service Initiated Connection Control* (SICC) interface.

Different configuration and deployment of this SICC interface in the network leave us with several different ways for the SICC switch to set up the required connections:

1 The SICC switch is a transit exchange. The connections are set up by signalling towards all participants, and tied together in the SICC switch.

2 All local exchanges (LE) are SICC switches, i.e. there is one SICC connection to every LE. Then the LE could set up a connection to its subscribers line, and signal towards all other end-points according to current NNI signalling, finally connecting them all together. This would give better network utilisation.

A further evolution of scenario BB-2 could be to have SICC interfaces on all switches in the network. This could then easily migrate to a fully compliant TINA Connection Management [9] that requires no signalling, but have direct contact for connection establishment with all switching resources in the network.

For narrowband/IN networks the group has focused on how to migrate the current IN architecture – ITU-T IN CS-1 [10] – towards TINA. The migration of IN is assumed not to happen in one large step, but rather in separated islands. This because of the large investments made in current IN and switching equipment, which are tied very close together. However, the functional entity SDF in IN is less close to the network, and also has no standard interior[1]. Consequently, these entities are candidates for migration. The object-oriented principles of TINA could be applied here to enhance the development and management of the SDF.

Scenario NB-2 proposes one more step by also developing the SCF according to TINA principles. This gives a unique environment for all IN functional entities except for the SSF, as the interface towards the SSF is assumed to still be according to the IN standard. The full TINA approach for narrowband/IN would require a change in the switching equipment, either by migrating the SSF towards TINA and introducing interaction between the SSF and the service objects (former SCF) through the DPE, or to take the full step by introducing bare cross-connects without any signalling and call handling capability.

As stated in the introduction, a fully compliant TINA architecture for narrowband is less likely because of the large investments in legacy systems, and the TINA requirements of major changes in most network resources for full compliance. However, new operators without any legacy systems might find simple switches and advanced and flexible software architectures like TINA interesting for management and provision of services, also in their narrowband network.

Service and network management aspects

The goal of this workgroup is to analyse the available architectures for management of network and services, and study a possible path of migration from these architectures towards the TINA Management Architecture.

The objective also implies an analysis of the benefits coming from the adoption of a TINA-based architecture.

The identification of a migration path towards the TINA architecture has been done according to the following steps:

1 Description of a reference architecture to be used in the provisioning of the MoD service. The architecture considered is to be the one exploiting management functions on existing systems.

2 Description of a set of relevant management scenarios (use cases) useful for the definition of management information to be exchanged among the architectural components.

3 Identification of the needed TMN [5] components for the "implementation" of the use cases and identification of reference points between those TMN components.

4 Identification of TINA components for the "implementation" of the use cases and identification of operational interfaces.

5 Identification of migration steps by mapping the identified reference points in the TMN architecture and the TINA interfaces.

The workgroup has made a comparison between TMN and TINA, and a first conclusion suggests that TMN can profit from TINA on the following two levels:

1 Modelling approach in order to reduce TMN complexity (especially on the computational model or in TMN terms on the functional model).

2 Distributed platform. TINA is based on a Distributed Processing Environment [7] (DPE) with distributed functionality, while TMN uses objects which have a fixed location. Nevertheless, there are already some trends in order to define generic mappings between Object Management Group (OMG) and TMN's Guidelines for the Definition of Managed Objects [6] (GDMO).

TINA can also profit from TMN. The following important areas are so far identified:

1 Network management (e.g. fault, performance (QoS))

[1] *Some work on this is taking place in IN CS-2, DFP.*

2 Inter-domain Management.

Steps in the migration towards TINA are being developed, including the aspects of interworking. During this work the group has identified a set of stakeholders for their application case, the Movie-on-Demand (MoD) service:

- MoD End User
- MoD Server System Provider
- MoD Service Gateway
- MoD Content Provider
- MoD Service Provider
- Network Provider.

These stakeholders and their interests were used to evaluate how TMN and TINA were able to cope with the different aspects of MoD management. The MoD management architecture was first described according to the TMN architecture. The same functionality deployed by the TMN systems was then represented according to the TINA concepts.

One important result of these analyses have so far been the identification of the need to have clear reference points in the TINA architecture, to be able to agree on inter-domain interaction (like the x-interface in TMN). The group's proposals for the MoD service have been given as input to the TINA-C, and work has been started there to establish the concept of domain types and reference points in TINA.

Information services

This workgroup has been focusing on identifying properties of the Internet and Information Services – such as the World Wide Web – that make it attractive. The motivation for doing this is later to study these services in a TINA context, to see how TINA is able to support these services, what the potential benefits will be, and to give some feedback to the TINA core team.

The study of the Internet has tried to identify features of the Internet information services that make the services interesting for relevant actors: user, service provider, network provide, and content provider.

The features in Table 2 show the features identified as important to the different actors of the Internet information services, which later will be evaluated for TINA.

Table 2 Important Information Services features for different actors

Actor: User	Actor: Service Provider
Access to repositories	Interaction with a Content Provider
Transparency	Browsing directory
Hypermedia	Security issues
Homogeneous interface	User access to services
Asynchronous Transfer Mode (ATM)	Advertising and finding services
Charging	
Screening	
Security	
Actor: Network Provider	**Actor: Content Provider**
Information services generate traffic	Relationship with Service Provider
Irregular traffic patterns, dynamic bandwidth	
Negotiated QoS	

P508 project plans

The project plans for 1996 are still under evaluation, but the main focus will be slightly shifted from migration and evolution towards interworking between legacy systems, the identified migration steps, and a fully TINA compliant platform. In the process of doing so some of the migration scenarios will be detailed further. Also, more effort will be given to promote the results towards standardisation bodies, ATM-Forum and TINA-C.

Future work

Telenor sees this project as an important link between the short term research on further development of the current IN architecture, and the longer term research which focuses on Open Distributed Processing (ODP) principles and TINA. This project, in connection with the work on IN/B-ISDN interaction in Eurescom P506, and the work on Mobility and IN in Eurescom P507 (see other article in this section), are all projects that investigate different development paths for the future of intelligent networks. These thoughts are probably shared by several of the partners in Eurescom, as all projects were highly attracted by most partners. The current partners in P508 are: CSELT (Italy, Project Leader), CNET/French Telecom, KPN Research (Netherlands), BT, Deutsche Telekom, TeleDenmark, Portugal Telecom, Telefonica (Spain), Broadcom (Eire), Telecom Finland, AFT (Finland), Telenor R&D.

References

1 Eurescom. *Project 508 : deliverable 1.* Heidelberg, 1995

2 Handegård, T. The TINA consortium. *Telektronikk,* 90(2), 74–80, 1994.

3 *An overview of the TINA Consortium work.* Red Bank, TINA-C, 1993. (TINA-C, Doc ref: TB_G.HR.001_1.0_93.)

4 *Overall concepts and principles of TINA.* Red Bank, TINA-C, 1993. (TINA-C, Doc ref: TB_MDC.018_1.0_94.)

5 ITU. *Principles for a telecommunications management network.* Geneva, 1995. (ITU-T, M.3010.)

6 ITU. *Information technology : open system interconnection : structure of management information : part 4: guidelines for the definition of managed objects (GDMO).* Geneva, January 1993. (ITU-T X.722, ISO/IEC DIS 10165-4.)

7 ITU. *Basic reference model of open distributed processing : part 1 : overview and guide to use.* Geneva, June 1993. (ISO/IEC 10746-2.2/ITU-T, Draft recommendation X.901.)

8 ITU. *Recommendations on B-ISDN.* Geneva, 1995. (ITU-T Q.2931.)

9 TINA-C. *Connection management specifications, draft.* Red Bank, March 1995. (Doc ref: TP_NAD.001_1.2_95.)

10 ITU. *Recommendations on IN: CS-1.* Geneva, 1993. (ITU-T Q.1211.)

Methods for dimensioning of Intelligent Networks

BY TERJE JENSEN

As the demand for services based on IN increases, the need to have appropriate tools to support the design of IN structures becomes more pronounced. Several of these services can imply different characteristics compared to the basic telephony service as seen from the network operator's point of view. Therefore, new methods for the relevant tools are requested.

Some steps and aspects to be considered when dimensioning IN structures are outlined.

1 Introduction

Intelligent Networks (IN) are being utilised on a larger scale by several operators. More services are expected to be based on IN, which, when accompanied with higher usage, results in increased load on the relevant equipment. An additional factor is that several of the services become more sophisticated compared to the basic telephony service. That is, the signalling and processing activities involved in handling a call can be several times higher than for the ordinary telephony call.

Faced with this evolution and the apparent lack of tools supporting the deployment of IN, it was decided to establish a project within EURESCOM[1]. This project was later named "Methods and Specifications for Tools to Dimension Intelligent Networks"[2]. As reflected in the name, the focus is placed on elaboration of general methods and specifications rather than development of the set of tools themselves.

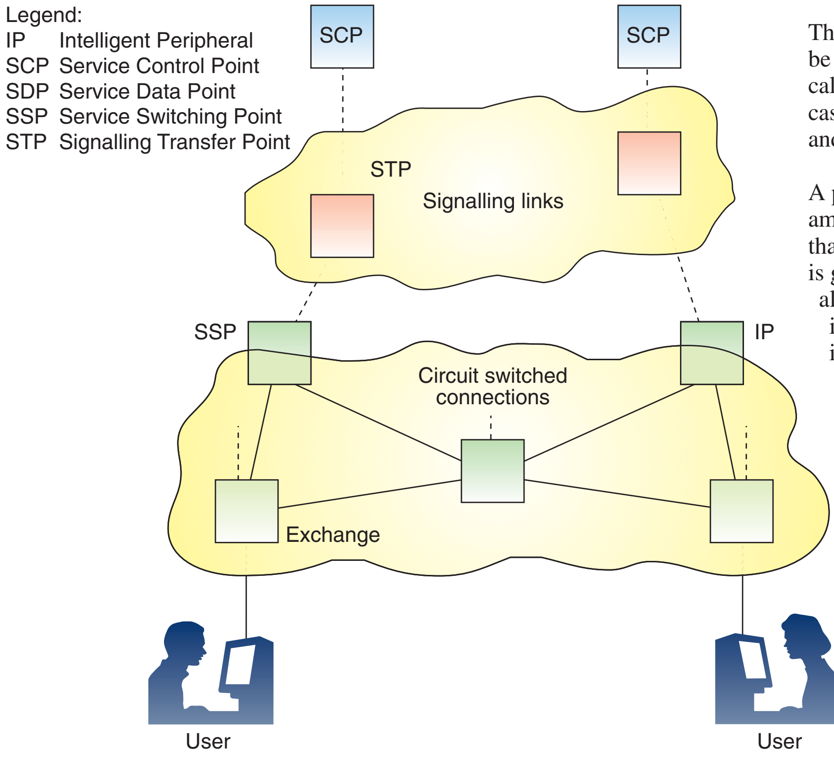

Legend:
IP Intelligent Peripheral
SCP Service Control Point
SDP Service Data Point
SSP Service Switching Point
STP Signalling Transfer Point
SCP
SCP
STP
Signalling links
SSP
IP
Circuit switched connections
Exchange
User
User

Figure 1 Physical elements and connections within the scope of the project. Several variants of the mapping of functional entities onto physical elements and connections between elements have been examined. These variants are not illustrated in this figure

Before the work could be started, the scope, steps and schedule were defined. In Figure 1 are illustrated the network elements and connections considered. The illustration does not show all possible connotations between elements and sub-networks. Different mapping of functional entities onto physical element are also possible, ref. [2]. Basically, the work was limited to Capability Set No. 1. As few results were available for the questions that arose, the dimensioning process itself also had to be described. Then, central questions for the technical planning of IN deployment were defined. Some of the activities carried out in order to provide answers for these questions are outlined in the following sections.

2 Dimensioning processes

In Figure 2 some of the steps in the dimensioning process are sketched. The main groups of input data can be categorised as

- characteristics of the service demand patterns
- service quality requirements
- characteristics of equipment/elements and the alternative network structures.

One major group of input data is the service demand. Not only the demand given as number of calls per second and holding times are to be given, but also the descriptions of the services themselves. In particular, for services based on IN, the operator/provider has flexibility when designing the services according to customers' needs. Therefore, more detailed service descriptions must be provided. This includes the composition of services by blocks which, again, specify the potential exchange of messages and processing steps involved.

The different call cases must also be considered. A call case can be defined by the destiny of the call, like successful/answered call, busy B-party, no answer, and so forth. These different cases may result in a different number and types of messages and processing tasks.

A parametric description of the service usage is defined as the amount of circuit switched, signalling and processing capacity that can be needed per service call of a given case. Usually, this is given with reference to the functional entities in order to allow for flexible mapping onto physical elements. For instance, the Virtual Private Network (VPN) service could include a number of features. Depending on these features, the number of messages and needed processing tasks could differ. For each call instance of the cases considered, the corresponding load on each relevant network component must be given. In case of VPN, a circuit switched connection to the SSP will be established. Some messages could be exchanged between SSP and SCP through the signalling network. Processing is needed in the involved entities. Then, a circuit switched connection from the SSP to the called user should be established. For some services the behaviour of the users should also be given. In par-

[1] *EURESCOM = European Institute for Research and Strategic Studies in Telecommunications GmbH.*

[2] *The project was defined in two phases: the specification aspects were added in the second phase.*

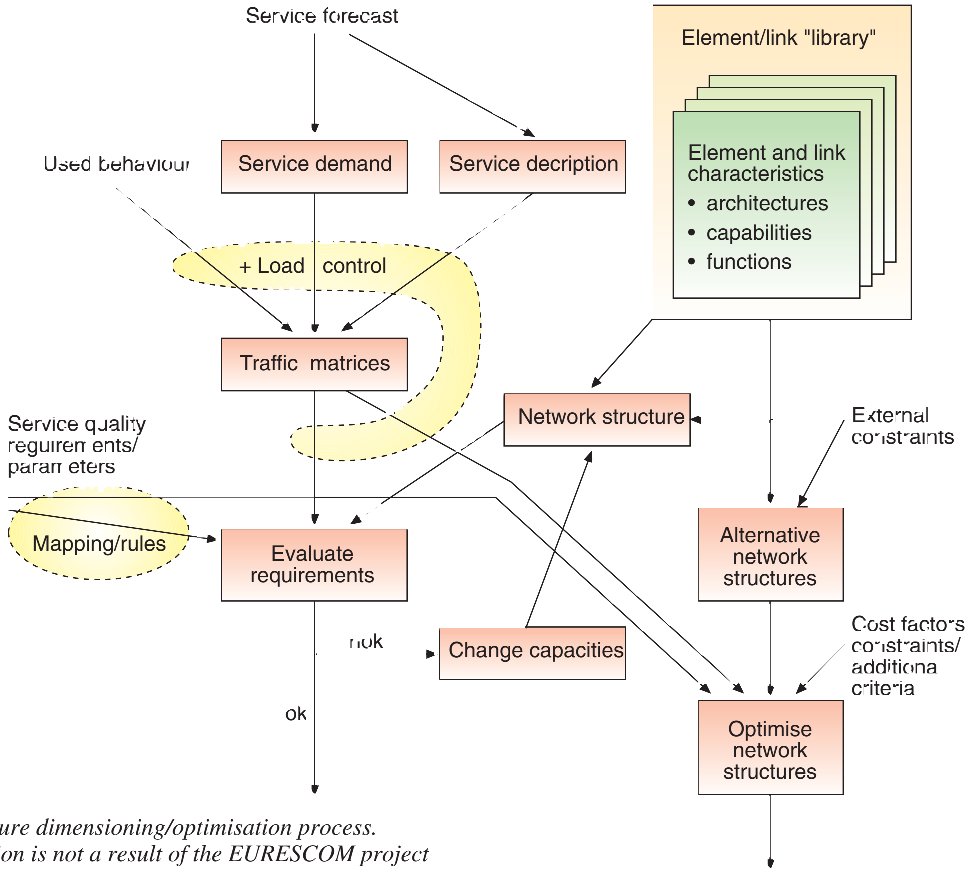


Figure 2 IN structure dimensioning/optimisation process. Note: this illustration is not a result of the EURESCOM project

ticular, this could be requested when non-stationary and high blocking situations are studied.

At this stage, traffic matrices can be established. Several sets of traffic matrices are relevant; those related to the circuit switched network and those related to the signalling network can be identified as two main groups. Although the telecommunications network is utilised for both IN-based and non-IN-based services, only the IN-related equipment is focused on during this process. However, non-IN-based services can also be characterised with additional traffic matrices. The matrices describe the amount of traffic that is requested from a source to a destination. What will be the sources and what will be destinations are chosen before establishing the matrices. Considering the VPN-service, the matrix of demands on circuit switched part could give the load from each local exchange towards the relevant SSP as well as the load from each SSP towards the local exchanges. For other services, like Televoting, only load from the local exchanges towards SSPs/IPs are relevant.

Although a number of cases and number of services can be identified, these are usually limited to a few variants. This is mainly done in order to restrict the number of variables (lower dimensionality). In addition, it could be an extensive activity to provide the requested input data if a lot of variants were kept. It might be requested to specify the demand for services in several intervals during the day. In particular, this is relevant when studying the combination of services. As an example, it is likely that VPN and Televoting have different demand patterns during the day with VPN having highest load in the working hours while Televoting may be mostly used outside the working hours.

3 Load and mechanisms

When the traffic demands are carried in the same network, the corresponding matrices could be combined. Several methods have been proposed for the operation of joining matrices. Which matrices to combine must also be decided. For instance, services having similar characteristics and utilising the same part of the network could be joined. As an additional factor, different sets of matrices could be established for the different intervals of the day. In case a multi-period dimensioning process is relevant, the matrices should be defined accordingly. For several cases, a single-period dimensioning process is sufficient which implies that the different periods could be considered by appropriate weighing of the service demand, that is, when combining the matrices.

For some situations the behaviour of the users should also be considered. For instance, when high blocking is experienced, call retries may result in considerable increased call rates. For services requiring much signalling and processing, the call rate could be a limiting factor. This depends on which phase of the service execution the call is blocked in.

In particular, for mass calling services overload control mechanisms must be introduced. Call gapping, e.g. ref. [1], has been described for use in IN Capability Set No. 1. This means that a fraction of the calls could be blocked in order to avoid that certain elements are placed in an overload condition. As an IN configuration could be looked upon as a centralised structure, for example the Service Control Point (SCP) might be in danger of getting overloaded. However, depending on the situation and how the services are implemented, the capacity of other elements could also be critical when mass calling services are active. The resulting demand patterns from applying such services depend on how they are utilised, for instance in a limited geographical area, viewers voting for a limited time, accepting the first caller only, and so forth. Televoting is an example of a mass calling service. This can be implemented in a number of ways. One possibility is to send a message from the SSP to the SCP for each call that is made. The counter can then be located in the SCP. Another implementation is to have counters in the SSPs and collect their values at certain instances. In the first case, the SCP may become heavily loaded when the service is activated.

Overload control mechanisms are usually present in most of the elements, although possibly designed for other purposes than the call gapping mechanism described for IN. The influence from these overload mechanisms may imply that the characteristics of the service demand are modified. The effects of these modifications depend on the level of the traffic load. For some of the services we can expect that the network is dimensioned to handle the load without activating the overload mechanisms. That is, these mechanisms are not considered during the dimensioning process. For other services, however, it may not be economically feasible to dimension the network to deal with the service demand. This can imply that an overload situation is expected to emerge when the service is utilised and that the effect from overload mechanisms must be considered. Examples of such services can the be VPN-service and Televoting service. For VPN we would usually prepare the network in a way to be able to handle the estimated demand with specified service quality (e.g. blocking). For Televoting, however, the demand could be very peaked (high demand in short intervals) making it impractical to apply the same set of values for the dimensioning criteria. In order to prevent the load from the Televoting service to influence too much on the handling of other services, special means could be introduced, like the call gapping mechanism.

4 Dimensioning of elements

The main groups of service quality variables are the blocking probability and the delay. Therefore, corresponding expressions for estimating these variables have to be established. For some cases, general approaches can be utilised, like establishing and analysing a queuing network in order to estimate the delay. In other situations, e.g. for the non-stationary characteristics of the traffic demand for a mass calling service, additional expressions could be requested. All the elements and connections considered for an IN structure must be handled.

In order to establish a "library" of element characteristics, the different elements and types of elements must be examined. One objective of this is to identify which element to apply for a given load intended for that type of element. For example, it is of interest to identify which type of Signalling Transfer Point (STP) to use for a given signalling load on that STP. As some functionality could be common for several of the elements, general capabilities of the functionality could be described. For instance, the functions related to a Signalling Point (SP) are present in the elements connected to the signalling network. Therefore, these functions could be analysed in view of the various ways of implementing them.

Most of the functional entities can be introduced in a stand alone or an integrated mode, see [2]. When the function is integrated with other functional entities, the common use of equipment must be considered. On the other hand, use of stand alone elements could imply an increase in the number of connections (circuit groups and signalling links). As it is difficult to say which solutions that are of more interest without considering the related costs, most of the alternative implementations should be considered.

All the element types present in the IN structure must be dealt with. That is, Service Switching Point (SSP), Intelligent Peripheral (IP), Service Control Point (SCP) and Service Data Point (SDP). In addition, other components, like STPs, exchanges, signalling links and circuit groups, could be included, see Figure 1.

During these examinations the hardware architecture of the elements must be taken into account. Different sets of capacity constraints could be the most pronounced ones for the different architectures. For instance, in centralised architectures the number of periphery processors/modules that a central processor can handle may be one constraint. This may not be stated for a distributed architecture. Considering an IP, the number of modules for circuit switched connections, signalling links, different devices and control units could be the outcome from the dimensioning activity. In addition, the software architectures must be considered. After mapping the software components onto the different processing elements, the requested capacity can be estimated. This is more relevant for delay calculations. Blocking calculations should also be performed when circuit switched connections are present in an element. In addition, the contribution to the service quality degradation should be estimated. Basically, maximal values for such contributions could be stated as requirements for dimensioning of the elements. However, as there could be trade-offs between the amount of degradation from each of the possible elements, different maximal values might be relevant.

The results from these activities would be descriptions of the different element types and how to find the relevant element to apply, fulfilling the stated requirements, when the load offered to the element is given.

5 IN structures

The structure of an IN includes all the elements involved in the IN related call handling. In addition, the connections between these elements and the distribution of traffic (routing) are described. The elements in the structure are taken from the "library" of elements described above. When identifying the possible structures, relevant equipment and external requirements on the network are to be considered. Two different purposes for analysing the IN structure are described in the following. These are given as end-points at the bottom of Figure 2. For the first one the topology (location of nodes and connections between nodes) is given and the capacity of nodes and connections are to be decided. In the second one, the location of nodes and connections are to be found as well as their capacities. In addition, a third purpose could be described, when both the topology and the capacities are fixed and it is to be examined whether or not an increase in the traffic load can be handled by the same structure.

When the topology is given, which capacities of the elements and connections to choose are found from the dimensioning rules established in the "library" for the corresponding equipment. However, for some cases the traffic distribution (routing) could be varied as well. Then some measures of the relationships between the capacities and the routing choices should be established as trade-offs between the capacity of which elements to modify can be present. An economic measure is a natural one, that is, a certain capacity of an element is associated with a cost for that element. Then, the capacities and routing is to be defined such that the cost of the network structure is as low as possible. In most cases, however, the routing rules are kept and the capacities are to be found. An economic measure would then be of less importance.

When dimensioning a network structure on longer terms, the topology as well as the capacities can be variables. Then, a suitable measure must be identified in order to compare the alternative structures. Again, the set of alternative structures are given by the components found in the "library". However, when analysing the structure, a more coarse model of each of the element candidates is usually introduced. For example, the modular capacity function given in a staircase form could be approximated by a linear function. The rationale for applying more coarse models is to limit the number of variables to be considered in this activity. Although it might be claimed that this makes the model more inaccurate, the accuracy of the other parts of the problem formulations must also be taken into account. For example, the predicted demand patterns for the services are given with a certain accuracy, the prices of the element types may not be absolute, and so forth.

When designing a network structure, the performance during a non-failure as well as the most common failure situations is verified. In particular, ensuring sufficient traffic handling capabilities during failure situations may become more pronounced as the dependency of telecommunication services and the competition increases. On the other hand, considering failure situations could result in higher modelling dimensionality. Again, a trade off between the modelling accuracy and the tractability of the resulting model could be identified.

The output from the dimensioning/optimisation step should be a structure fulfilling the stated constraints. For the optimisation the resulting structure should be close to the optimal one considering the objective measure that has been used. When the sum of the cost of the relevant equipment is used as the objective measure, this should be as low as possible.

6 Service quality variables

Here, the term service quality variables includes the sets of variables used for characterising the quality of services and the performance of networks. Such variables could be defined as seen from the user or for a portion of the network. During the dimensioning process, both view points are considered, like end-to-end blocking and the delay for a message passing an STP. However, during the optimisation, these requirements are often decomposed onto the various sections of the network. This can be supplemented by how many sections of the different types that can be used in sequence.

Target values for the service quality variables are usually decided by the environment of the operator, like by international standardisation bodies. In addition, the definition of service quality variables is also to be treated in these organisations.

Service quality variables should be measurable. Therefore, appropriate measurement schemes must be described. Although some specific aspects for IN-based services could be found, the measurements should utilise the common mechanisms already present in several of the elements. In that way, procedures carried out by the operator for other purposes as well could be applied. Some specific post processing could be needed in order to find estimates for the relevant variables. For a few of the variables, specific measurements are foreseen as well.

7 Related activities and further areas

Several of the central methods elaborated in the EURESCOM project will be implemented in prototypes. The rationale for this is to evaluate the methods themselves (accuracy, computer times, etc.) and to gain some experience in designing IN structures by applying the implemented methods on some cases. Naturally, during this process, some of the steps in the methods could be refined based on the observations made.

Parts of the results have been contributed to ITU-T, in particular for Study Group 2 (Network operation). Some of the topics dealt with in that group, related to the areas studied in the EURESCOM project, are performance of IN and signalling systems (Question 12/2: traffic for IN and No. 7). Measurement procedures are also treated in that Study Group. The contributions from the EURESCOM project seem to be met with a certain interest as few similar results are known. Currently, there are no activities on these topics in ETSI.

Representatives from Telenor have been involved in several of the activities from the start of the EURESCOM project. One reason for this is to gain insight into the different aspects of the dimensioning/optimisation of IN structures. In addition to the experience gained, a number of methods for dimensioning of such structures have been elaborated. Knowing the details as well as the assumptions these are based on, alleviates the process of establishing a tool for designing IN structures. Schemes and procedures for estimating the resulting performance and quality for services utilising the equipment should also be applied. This is of particular interest when introducing more services and customers, like during the service creation process. Then, the quality indicators of all the services should be considered during the design and definition of the additional service demands.

It is expected that further requirements on "openness", future capability sets and more general distributed systems will emerge. In this view, the defined procedures may work as a fundament for studying more enhanced problems. This is supported by the knowledge about suitable models and analyses that have been achieved.

References

1 ITU. *Distributed functional plane architecture for intelligent network CS-1*. Geneva, 1995. (ITU-T draft rec. Q.1214.)

2 ITU. *Physical plane for intelligent network CS-1*. Geneva, 1993. (ITU-T rec. Q.1215.)

Mobility applications integration in Intelligent Networks

BY GEIR OLAV LAURITZEN

Introduction

Intelligent Networks (IN) is used in the fixed networks today to enable a faster and easier introduction of new services and to achieve a more efficient use of network resources. IN is also very well suited for implementing personal mobility. Personal mobility enables the user to access telecommunications services from a variety of terminals and network access points. In fixed networks, personal mobility has been realised through the service concept Universal Personal Telecommunications (UPT). In Norway UPT phase 1 is already implemented, and it is a growing service in the Norwegian telecommunication market.

One step forward from UPT is to provide support of terminal mobility on the IN platform. Terminal mobility enables the user to carry his terminal with him, and the network keeps track of his position in order to deliver incoming calls to him while on the move. This functionality is already known from mobile networks like GSM, but work is currently being done to integrate terminal mobility also in the fixed networks. This service is called *Cordless Terminal Mobility (CTM)*. In ETSI the CTM concept is currently being standardised based on IN on the network side and *DECT (Digital European Cordless Telecommunications)* or *CT-2 (2nd generation Cordless Telephone)* on the radio access side. An integration of terminal mobility features in IN will be advantageous. IN will then provide a common platform for mobility and service handling. The service data and mobility data can be integrated, and a more efficient data storing and handling can be achieved.

The *EURESCOM project P507 "Mobility Applications Integration in IN"* covers activities aiming to support mobile services on future IN architectures by considering the CTM service as a basis for the short term scenarios. The UMTS (Universal Mobile Telecommunications System) concept is studied as a target scenario for evolutionary paths from existing telecommunications systems towards the 3rd generation mobile systems. In addition, data modelling aspects which deal with the development of an information model are studied. Telenor is not involved in the information model development and consequently, the focus for this article is mainly concerned about the work on CTM and evolution towards UMTS.

The P507 project is defined as an extension of EURESCOM P230 project *"Enabling pan-European services by co-operation between PNO's IN platforms"*. P230 was as the title indicates, mainly focusing on internetworking at the IN-level across network borders, but was also dealing with mobility aspects based on IN. The P230 work on mobility was among other things concerned about DECT radio access enhanced with IN mobility control, and the basis for the current CTM (Cordless Terminal Mobility) functional architecture was developed by the P230 project.

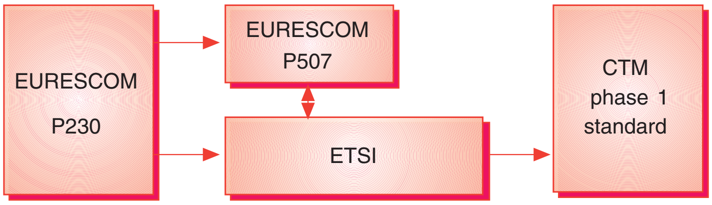


Figure 1 Development of CTM phase 1

Cordless Terminal Mobility (CTM)

The technical work on CTM starts with the EURESCOM P230 project, where the basic functionality for CTM was developed. Much of the results obtained in P230 was adopted by ETSI. The main focal point for CTM studies in ETSI is STC NA6, currently in charge of developing standards for phase 1 of the CTM service. This phase aims to provide intra network CTM support for IN-structured public networks and support for CTM internetworking between an IN-structure public network and Private Telecommunication Networks (PTNs). ETSI STC BTC1 also studies internetworking between public networks and PTNs, but in contrast to NA6 CTM focuses on the PTN operator rather than the public network operator (PNO). One example of where public-private internetworking can be useful is when the customer is visiting a hotel that has its own private CTM network implemented. The customer should then be able to use his own cordless terminal to access this network, and be charged to his account in the home network.

The work in P507 is concerned with a global CTM service, the work within this project is therefore focused on internetworking. Based on the work in NA6 CTM and BTC1, two main goals have been defined for the CTM studies in P507 (see also Figure 2).

1 To aid the work on CTM internetworking between an IN-structured public network and a PTN in NA6 CTM and BTC1

2 To enhance the CTM service to also support internetworking between IN structured public networks. This is useful for e.g. business customers travelling abroad.

To reach the goals listed above, a technical approach described hereafter is adopted.

1 Feasible scenarios for internetworking between public-public or public-private networks is identified. The difference between the scenarios is the way internetworking is done over the network boundaries, which means that a scenario is defined on the basis of which functional entities that are talking to each other.

2 In the cases where more than one scenario is defined, generic information flows describing the information exchanged between the functional entities on each side of the network border, are developed. These information flows are independent of the particular scenario, and are describing the message exchange necessary for each of the basic features: location registration, terminal authentication, incoming call and outgoing call.

3 The next step is to map the generic information flows onto the specific scenarios identified in pt. 1, and develop scenario specific information flows.

4 Finally, an evaluation of the different scenarios is performed to identify requirements and limitations on each scenario.

Evolution towards UMTS

Evolution from existing telecommunication systems towards UMTS has become a more important issue these days. Operators have made great investments in their existing networks, and reuse of equipment and technologies for UMTS is therefore of interest. There are several interest groups that will influence on the evolution paths towards 3rd generation mobile telecommunication systems. This can be observed in their work in the current standardisation process. The most obvious interest groups and their views of evolution are summarised below.

- *Fixed Operators* would like to see an evolution based on an extension of the fixed network. This implies that an integration of mobility features in the already existing IN platform is preferable. The development of CTM can be considered as a part of this evolution.
- *Mobile Operators* have already the mobility functionality implemented in their GSM networks, and foresee an evolution path from GSM by extending the service repertoire. The introduction of IN services is therefore the next step foreseen. The CAMEL[1] feature is currently being standardised in ETSI based on the current IN standard, and will enable the customer to bring his customised services with him while roaming into other GSM networks.
- *Manufacturers'* interest is to keep most of the functionality of the system dependent on their products. The operator will then be forced to continue buying equipment from the vendor who has developed the equipment he is currently using, and therefore be dependent on this vendor's coming products and upgrades. This means that the manufacturers want the standards to give them freedom to implement their own solutions which cannot necessarily be replaced with another vendor's equipment.
- *Users* would require a system that will give some benefits compared to the systems they are already using. This could e.g. be an improvement of the quality of the existing services or an extended set of services. The need for broadband/multimedia services are increasing, and will most likely be a key factor for the success of UMTS.
- *Regulators* will make sure that a fair competition can take place in the telecommunications market after the re-regulation in 1998. Of special importance is that the coming systems provide capabilities for 3rd party service providers so that a competition can take place in the service market.

P507 is developing evolutionary scenarios that are using the following approach: Initial situations are identified by studying today's technologies and systems. From this initial situation evolutionary steps are defined preferably based on foreseen triggering events like e.g. milestones in relevant standardisation bodies.

Considering the systems and technologies existing today there are several possible candidates for migration/evolution towards UMTS.

- The *fixed network* is one major candidate using the IN-technology to provide services and mobility. Three parallel evolution components can be seen in the context of the fixed networks. The evolution of backbone networks (N-ISDN, B-ISDN) and the evolution of IN are two of these components, and they will be dependent on each other. It is important to have the necessary interactions when developing the corresponding standards in order to create a harmonised system e.g. with respect to mobility features like handover. Besides this, we have the development of radio interfaces, and it is assumed that there will be a revolution on the radio interface for UMTS.
- *GSM networks* can evolve by integrating IN to support customised services. In contrast to the fixed networks the GSM network already has the functionality for mobility management, and the first evolution step should therefore be concerned about the service aspects. Further steps can be an integration of mobility- and service-control into a common platform, and later introduce broadband capabilities by integration with B-ISDN. A new radio interface is also required to support the broadband services.
- A *Cable TV network* is also looked upon as a possible candidate for evolution towards a 3rd generation telecommunication system. The cable TV operators are already providing

[1] *Customised Applications for Mobile networks Enhanced Logic.*

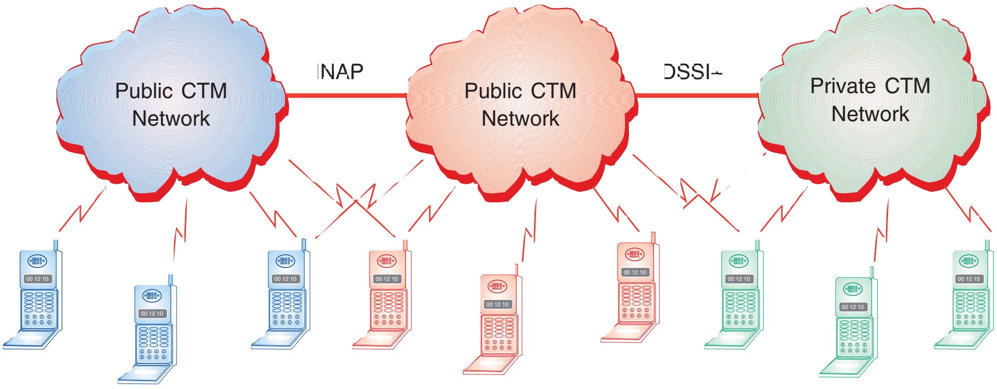


Figure 2 Internetworking between public networks and between private and public networks

broadband services to their customers, for the time being through broadcasting, but there will soon be a "claim" for Video on Demand services. This will require two-way communication paths, and this point can easily lead to provision of telephony services.

- *Satellite networks* can provide global coverage in the initial phase and are therefore considered as a candidate for migration towards 3rd generation mobile systems. This evolution path starts by providing satellite coverage and will be extended with a terrestrial infrastructure in a later phase.
- A roll-out of *Stand Alone UMTS* from scratch is not really an evolutionary scenario but must be considered because it is a relevant choice for new operators on the market. Such operators will not be restricted by investments in second generation technology, and might invest in the 'state of the art' at that point in time.

Information Model Development

P507 is also developing an information model which is based on object oriented methodology. The model is general but will be used in particular to analyse the CTM architectures developed.

The ETSI CTM-service is an effective means of connecting advanced cordless systems such as DECT or CT-2 into telephone networks through Intelligent Networks. The problem in the development of the CTM-service is that as new features are added to provide a more versatile service, the service becomes very complex. Understanding the service and the behaviour of the service as a whole thus becomes more and more difficult and time consuming. To help this situation and to avoid difficulties in the future, an object model of the CTM-service is defined.

In object modelling the real world items and phenomena are described as objects, relations between objects and dynamic behaviour of objects. The object model describes the general and detailed structure of objects in a system. Modelling can be done from different viewpoints. Static structure, object identities, object relations and attributes are described by the object model. Dynamic behaviour of all or a specified group of objects is described by the dynamic model. Finally, the functional model describes the information transforms inside the system and across the system boundaries.

Summary

The project was initiated in January 1995 and will be finished in December 1995, and two deliverables will be produced. Deliverable 1 contains intermediate results, and Deliverable 2 will contain the finalised results obtained in the project. For the time being it is not clear how much of this material will be publicly available.

The partners in P507 are: CSELT (Italy, Project leader), France Telecom/CNET, British Telecom, Telecom Ireland, KPN Research (Netherlands), Telefonica I+D (Spain), Deutsche Telekom, OTE (Greece), Telecom Finland, ATC Finland, Telenor R&D.

Reference

1 EURESCOM. *Project 507 : deliverable 1.* Heidelberg, 1995.

EURESCOM P408 "Pan-European TMN – experiments and field trial support"

BY PÅL KRISTIANSEN

The objective of this article is to present an overview of the ongoing EURESCOM project P408. P408 is the first experimental project within EURESCOM and involves several parallel activities. The total manpower allocated to the project is at the moment 540 man months over a period of 33 months. P408 is expected to finish in December 1996.

Background

EURESCOM has, since the start in 1991, launched several projects in the TMN area addressing the problem of management of pan-European networks and services. These projects had, until 1994, produced theoretical results such as guidelines, models and specifications. In order to increase their value, these results should be tested, validated and enriched by experiments. This requirement lead to the launch of P408 in March 1994.

The objectives of the project P408 are

- to provide the establishment of the PET-Lab *(Pan-European TMN Lab)* as an experimental system for pan-European TMN and to provide support for experimentation and testing of inter-domain management via the X-interface [1]
- to complete and refine specifications of the X-interface for an implementation of inter-domain management of ATM/SDH networks
- to coordinate the specification, design and test of HMI *(Human Machine Interface)* in this project in order to run simultaneous experiments on the X-interface in the various laboratory sites
- to provide support and specifications of security functionality in order to protect the interactions between individual TMN systems according to the joint/cooperative management models
- to direct and organise the work to complete the X-interface for its implementation.

Table 1 lists the 13 contractors involved in the project.

Project overview

The project is divided into two, part A and part B, which are mainly running in parallel. Figure 1 shows the main activities in each part.

Part A is concerned with the establishment of PET-Lab and with the conduction of experiments within the established environment. Part B is concerned with refining the TMN X-interface specifications for SDH and ATM management. The relation between part A and part B is that part A focuses on conducting experiments defined by part B and delivering test results to part B. Activities on Security and HMI are also included in part B and are the only activities in which Telenor is involved. The rest of this article will give a short overview of each of the activities denoted above.

Table 1 List of P408 contractors

P408 contractors
Deutsche Telekom AG
Telecom Finland Ltd.
The Association of Telephone Companies in Finland
Telecom Eireann
BT
Portugal Telecom SA
France Télécom
Tele Danmark A/S
Telenor AS
Telia AB
CSELT
Belgacom
Telefonica

Establishment of PET-Lab

The first step was to conduct a feasibility study to investigate and define the requirements for the set-up and operation of PET-Lab. Important issues were addressed, such as the definition of a general PET-Lab structure, the schedule for setting up and running PET-Lab, a selection of those TMN laboratories which would form the initial core of PET-Lab including requirements for these laboratories.

The initial core of PET-Lab consists of eight TMN laboratories which are divided into two separate groups, one SDH-group and one ATM-group. Figure 2 points out the eight lab-locations. Additional labs are planned to join the two groups in 1996.

Within a group, each TMN lab is interconnected with the other three TMN labs in the group. The DCN *(Data Communication Network)* infrastructure between the labs is based on X.25 PSPDN switched connections with an access rate of 9.6 kbit/s. The protocol profile chosen for the X-interface is QB1 (TP0,2

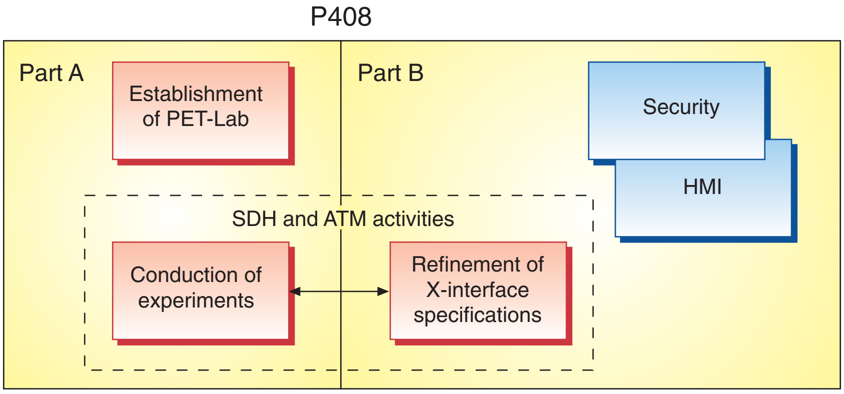


Figure 1 Project overview

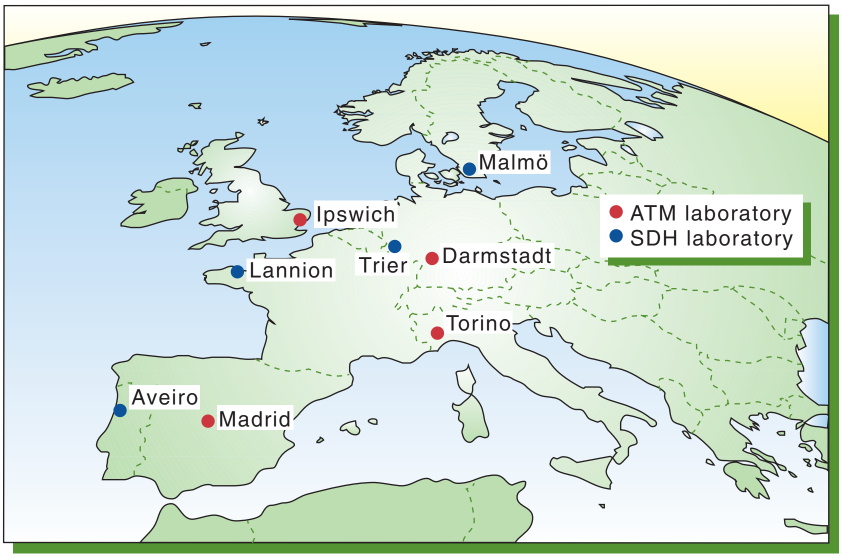


Figure 2 Initial core of PET-Lab

transport layer over X.25), according to ITU-T Q.811 [2] and Q.812 [3] recommendations. Each TMN lab will be connected to a subnetwork, either consisting of real NEs *(Network Elements)* in the laboratory, real NEs in a network or a simulator.

The various lab-owners are using management platforms from different vendors such as Bull, Ericsson, Siemens, HP and AT&T to provide the necessary runtime environment. In other words, PET-Lab provides a true distributed multi-vendor environment.

When it comes to interoperability between the various platforms, experience has shown that different implementations are not always compatible, even when they are based on the same standard. An important part of the PET-Lab establishment phase has therefore been to ensure good interworking between the various protocol stack implementations. In this relation IOP-tests *(Interoperability tests)* have been defined to verify good interworking on CMISE *(Common Management Information Service Element)* level between the labs. The IOP-tests were completed second quarter 1995.

Once operative, PET-Lab is a multi-site facility with the objective to provide support for experimenting and testing of inter-domain management via the TMN X-interface. The PET-Lab itself is not devoted to a particular technology, but provides the means to test and experiment in different areas. Within P408, though, the planned experiments are focusing on SDH and ATM.

SDH and ATM activities

The work for SDH management and ATM management is organised as separate activities. The objectives, though, are very similar:

- To complete and refine specifications for the X-interface
- To define test suites for the validation of the specifications
- To implement and run the tests in PET-Lab
- To analyse the test results and revise the specifications.

ATM and SDH were selected because of the key relevance of these technologies in future networks. Another thing was the presence, in the European arena, of two potential “clients” of the management specifications; METRAN *(Managed European TRAnsmission Network)* for SDH and the ATM Pilot for ATM.

For pan-European ATM management two types of X-interfaces are addressed within P408, namely

- The “Xcoop” management interface, which is defined as the interface between two Public Management Systems belonging to different Public Network Operators (PNOs). A Public Management System is in this case used for managing a public ATM network.
- The “Xuser” management interface, which is defined as the interface between a Public Management System and a Private Management System. A Private Management System is in this case used for managing an ATM CPN *(Customer Premises Network)*.

For pan-European SDH management, following the scope of METRAN, only the “Xcoop” interface is relevant.

The work for SDH/ATM can be divided into five basic activities, as shown in Figure 3. The first activity is concerned with producing the X-interface specifications that will be used as a basis for implementation and testing. The four other activities are related to the implementation and testing of these specifications.

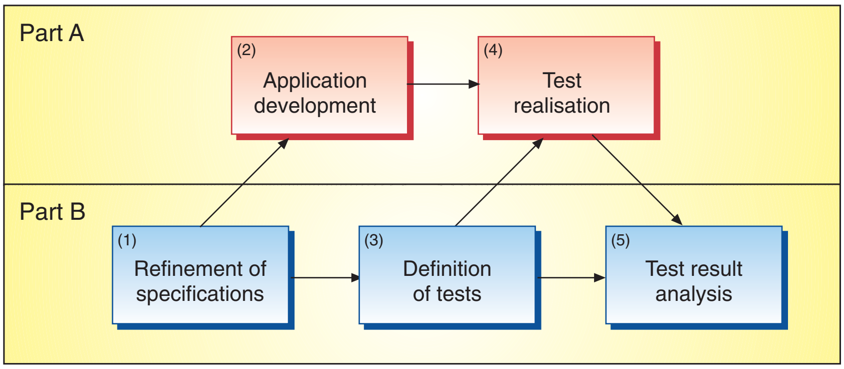


Figure 3 Basic activities for SDH and ATM

Refinement of specifications

To make the amount of work feasible within P408, it was necessary to focus on a small subset of management services.

For SDH management, the scope covers Fault Management and Path Provisioning. The aim is to produce results that will be applicable to METRAN, and more specifically the target METRAN solution where SDH network management will be fully automated between peer TMNs. The work is influenced by METRAN requirements and is using results from two previous

EURESCOM projects, P109 *("Management of Pan-European Transmission Networks")* and P107 *("METRAN Technical Network Management")*.

For ATM management, the scope covers Fault Management for the Xcoop interface and Subscription Management, Virtual Path Provisioning and Fault Indication for the Xuser interface. The results should be applicable for a future pan-European ATM environment supporting fully automated management. The work is using results from EURESCOM P105 *("Pan-European ATM-studies")* and from the ATM-pilot consortium.

The X-interface specifications are produced through the use of Ensembles. The Ensemble concept is defined by the Network Management Forum [4] and provides a framework for producing a complete description of a management service across a management interface. Briefly described, an Ensemble defines the resources to be managed, the management capabilities, the management information model and conformance requirements for the Ensemble. The production of Ensembles was finished during second quarter of 1995.

Implementation and testing

Activity (2) is concerned with the implementation of prototype applications based on the Ensemble specifications. A fundamental project principle is that each PNO providing a TMN lab is responsible for the development, reliability and integrity of its own applications within the lab.

The objective of activity (3) is to define test suites for validating the Ensemble specifications. The goal is to produce a set of Abstract Test Suites (ATSs): one for the SDH Xcoop interface, one for the ATM Xcoop interface and one for the ATM Xuser interface. In order to formalise the testing language within P408 it was decided to write the ATSs in TTCN *(Tree and Tabular Combined Notation)* [5] using methodology guidelines developed by EURESCOM project P201 [6] on TMN testing. The final ATSs are planned for the third quarter of 1995.

Activity (4) will convert the ATSs into executable test suites for the various implementations within each of the TMN labs. The tests will then be run both by internal staff to the TMN lab and by visiting staff from other PNOs. The testing activity is scheduled between third quarter of 1995 and second quarter of 1996.

Activity (5) will analyse the test results and use this to revise the X-interface specifications developed by activity (1). The final results will be submitted in two Deliverables, one for the SDH specifications and one for the ATM specifications.

Security

The activity on security provides a total effort of 47 man months. The target objectives are

- To specify a long term security solution that will facilitate secure interactions between TMN systems participating in pan-European cooperative management over the X-interface
- To validate and test parts of these specifications through implementation and testing of selected security features in PET-Lab.

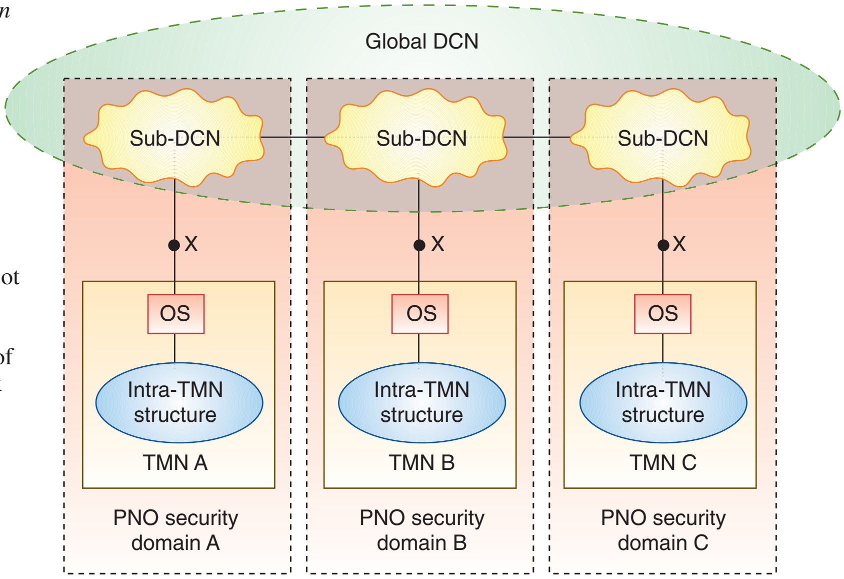


Figure 4 Pan-European TMN infrastructure from a security viewpoint

The main focus is to produce specifications that can be used as input to METRAN and/or a similar ATM-effort for pan-European network management. In addition, the specifications should find their value as input to standardisation activities addressing TMN security. One obvious target would be ETSI NA4 where a security rapporteur group was established in March 95 with the objective to produce European Technical Standards for TMN security.

Scope of specifications

Figure 4 shows a simplified pan-European TMN infrastructure as perceived from a security viewpoint. The figure introduces the term security domain. Briefly described, a PNO security domain will contain all elements being under that PNO's security jurisdiction. A general assumption is that every PNO is autonomous when it comes to defining its own security policy. This policy will apply in every TMN being part of that PNO's security domain. The figure illustrates a TMN as being a subset of a PNO security domain, but in many cases the two will coincide.

OSs *(Operations Systems)* belonging to different TMNs are interconnected end-to-end through a global DCN. The global DCN itself might consist of network resources within the jurisdiction of different PNOs. This is illustrated in the figure by letting subnets (sub-DCNs) of the global DCN be part of different PNO's security domain.

Inter-TMN interactions require cooperation across security domain borders. In order to achieve interoperability it is necessary that the PNOs reach common agreements for how to protect the interactions. These agreements will be part of a common security policy, normally referred to as a "secure interaction policy".

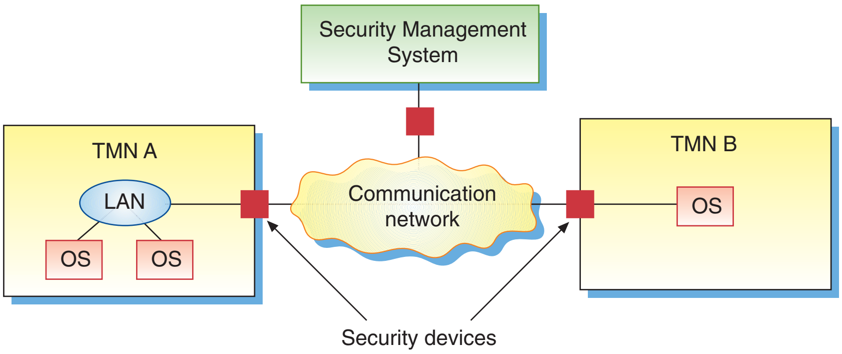


Figure 5 Principle of security front-ending

The scope of the long term security solution has been defined as to specify those parts of the “secure interaction policy” that are required to achieve security interoperability between TMN systems. The specifications will include recommendations for

- A common set of security services and security mechanisms
- Security placement
- Protocols for end-to-end exchange of security information
- A common representation (syntax) of the information to be exchanged
- Security management capabilities required for managing the security services and mechanisms
- The use of TTP *(Trusted Third Party)* services.

The work is partly based on security requirements for METRAN management which has been developed by the EURESCOM projects P107/P110 [7]. The final specifications will be documented in a separate Deliverable.

Implementation and testing

The activities concerning implementation and testing of security features in PET-Lab are performed at two stages.

The first stage is performed in parallel with the long term specification work. A subset of security features that might be applicable for the long term solution has been selected and the idea is, by the first quarter of 1996, to have implemented and tested out this subset, including management, in PET-Lab. The results will be used as input for the long term specification work. For practical reasons, the decision was to test the applicability of security front-ending.

Figure 5 illustrates the principle of security front-ending. The security features are integrated in stand-alone hardware devices which are front-ended with respect to the end-systems. In the implemented solution the devices are managed from a single remote management centre. The communication flow is protected during transfer between two security devices. One advantage of security front-ending is that several OSs may share the same security device, as shown for TMN A in Figure 5. Another advantage is that the security features provided will be totally transparent to the end-systems. The short time frame has made it necessary to use already available security products. Rather than buying new equipment on the market, Telenor has offered the project to rent the necessary X.25 front-end security devices including management. The result from this activity will be documented in a separate Deliverable.

The second stage of implementation will take place when the long term specifications have been completed. Some of the specified security features will be selected for implementation and testing in PET-Lab. According to plan, this activity will start the second quarter of 1996.

HMI

The work on HMI is a very limited activity, providing a total effort of 20 man months. The main objective is to coordinate the specification, design and evaluation of HMI for the ATM and SDH management applications that are developed for PET-Lab.

The activity is divided into two. The first part is concerned with producing a set of general guidelines meant to support the designers of user interfaces for the management applications. Topics like harmonisation of language and terminology, visualisation techniques, colours and symbols, are addressed. This part was completed second quarter of 1995. The results have been documented in a Deliverable.

The second part of this activity will seek to benefit from the practical implementations performed within PET-Lab. The idea is to evaluate the final PET-Lab applications against the guidelines produced. This way it should be possible to identify what could be improved in the original guidelines and to identify what could be improved in the implemented user interfaces.

The output of the HMI effort will, in addition to supporting the SDH and ATM experiments within P408, be used as input to other EURESCOM projects such as P414 (“TMN Guidelines”) which has a subtask producing user interface guidelines for TMN applications. Future standardisation activities on TMN HMI might also benefit from this work.

Future use of PET-Lab

Before P408 started, all EURESCOM projects have been of a theoretical nature (analysis, evaluation, specification, fact finding studies). The whole organisational framework was established with this theoretical work in mind. It was not obvious that an international testbed like PET-Lab could be established in the same way as other projects, e.g. by exchanging all the necessary information. With the success of P408 it has been demonstrated that a practically oriented project is really feasible under the conditions of the EURESCOM treaties and by using the existing framework. In other words, EURESCOM will have proven itself to be a well suited candidate for pan-European experiments in the future.

At the moment, there are strong indications that PET-Lab will survive its scheduled end in December 1996 as a pan-European TMN testbed. As a concrete example there is a proposal to extend P408 in order to continue the X-interface validation and verification for SDH for an extended set of management services. Similarly, it is also quite probable that proposals will be forwarded proposals to continue the work on the X-interface for ATM. Extended security experiments might also be suggested.

As mentioned earlier, PET-Lab is not devoted to a particular technology, i.e. the facility can be used to test inter-domain management between TMNs in general. The SDH and ATM related experiments within P408 are only the first step. There are EURESCOM shareholders who would like to make the next step towards a testbed for extended TMN tests. Possible extensions might be

- additional managed networks (e.g. ATM over SDH)
- additional managed services (e.g. charging, billing and accounting)
- new technology (e.g. CMIP *(Common Management Information Protocol)* over SDH ECC *(Embedded Control Channels))*
- the use of real network elements instead of simulators.

References

1 ITU. *Principles for a telecommunications management network.* (ITU-T Recommendation M.3010). Geneva, 1995.

2 ITU. *Lower layer protocol profiles for the Q3 interface.* (ITU-T Recommendation Q.811). Geneva.

3 ITU. *Upper layer protocol profiles for the Q3 interface.* (ITU-T Recommendation Q.812). Geneva.

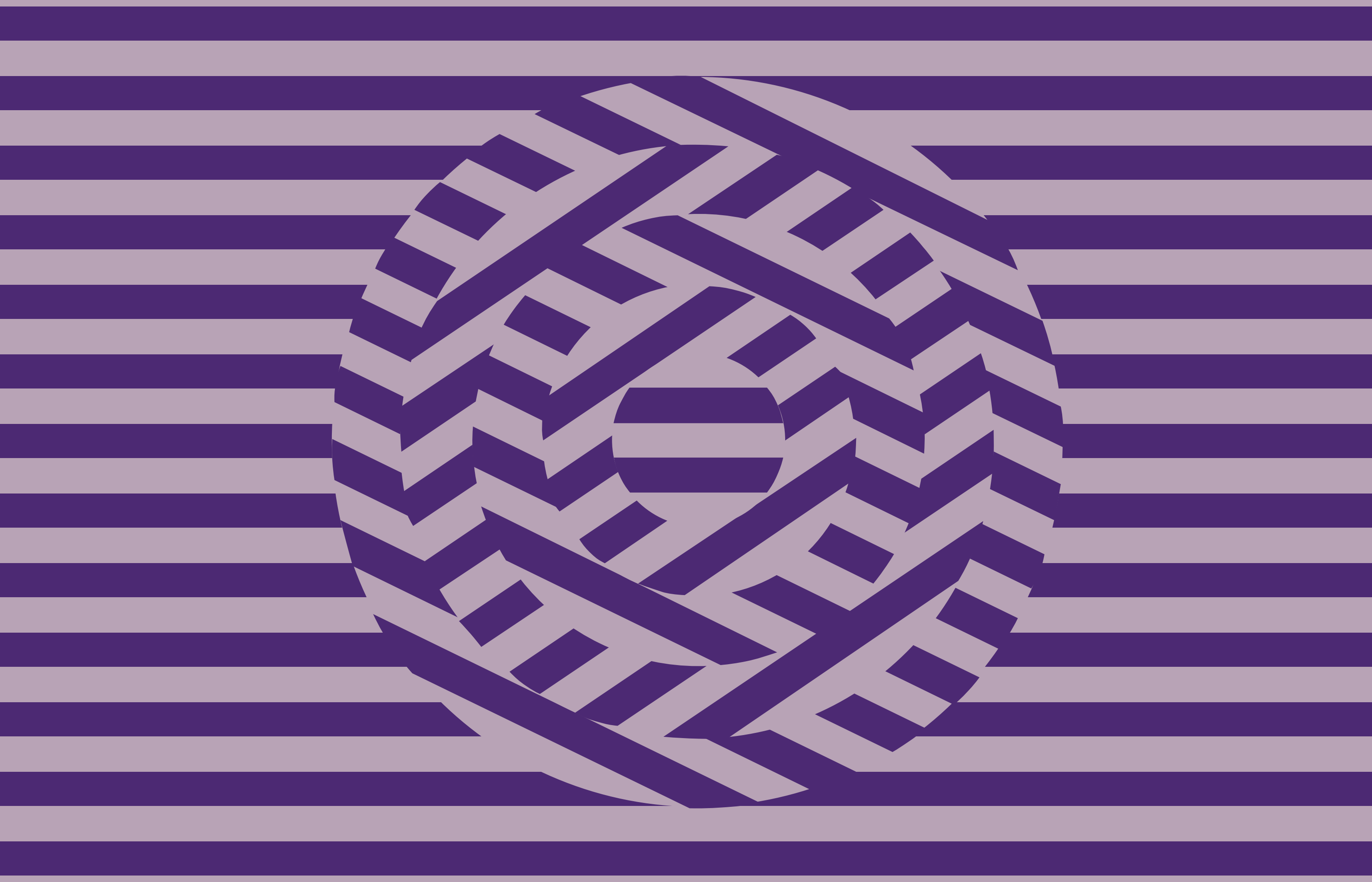
4 NMF. *The ensemble concepts and format.* Network Management Forum, 1992. (Forum 025, issue 1.0.)

5 ISO/IEC. *Conformance testing methodology and framework.* (ISO/IEC 9646, part 1–4).

6 EURESCOM. *IN&TMN service testing.* Heidelberg. (EURESCOM P201, Deliverables D2, D3 vol. 2, D4 vol. 1/vol. 3.)

7 EURESCOM. *Security of METRAN management.* Heidelberg, 1994. (EURESCOM P107/P110, E-METRAN Task F, Subtask F8, vol. 1&2.)

Kaleidoscope



90 years of Televerket's technical journal – an overview

BY HENRIK JØRGENSEN

Early beginnings, 1904 – 1909

The first issue of *TEKNISKE MEDDELELSER fra Telegrafstyrelsen* (TECHNICAL INFORMATION from the Telegraph Administration) appeared in April 1904, and was introduced by a short statement of objectives:

> "The Telegraph Administration herewith produce the first issue of the news sheet *Technical Information*. It is intended that this will impart knowledge, to all personnel, of the State Telegraph technical installations, and will, in short, supply the more important news in the areas of telegraphic and telephonic technology. As a rule, the news sheet will appear once a month, at the same time as the monthly circular."

The first *Technical Information* was entitled 'On the protection of telephone cables from power lines', and extended to two issues. Even this early in its history there was introduced the editorial mark of the journal that it should contain both short and long articles concerning technical matters, and small announcements on technical news.

TEKNISKE MEDDELELSER

FRA

TELEGRAFSTYRELSEN.

Nr. 1	April	1904

[illegible]

Om beskyttelse af svagstrømsledninger mod stærkstrømme.

[illegible]

Facsimile of the front page of the first issue of Technical Information from the Telegraph Administration, *1904*

This first period is characterised by an anonymous strongly official production style, which occasionally has the appearance of technical decrees. It is only in the next stage of this organ's life that ascribed articles appear.

Eight issues appear in 1904, and six in 1905. Thereafter the production flow ebbs away, with a decline in frequency to the point of only one issue in each of 1908 and 1909. Thereafter the official mouthpiece is struck dumb for six years.

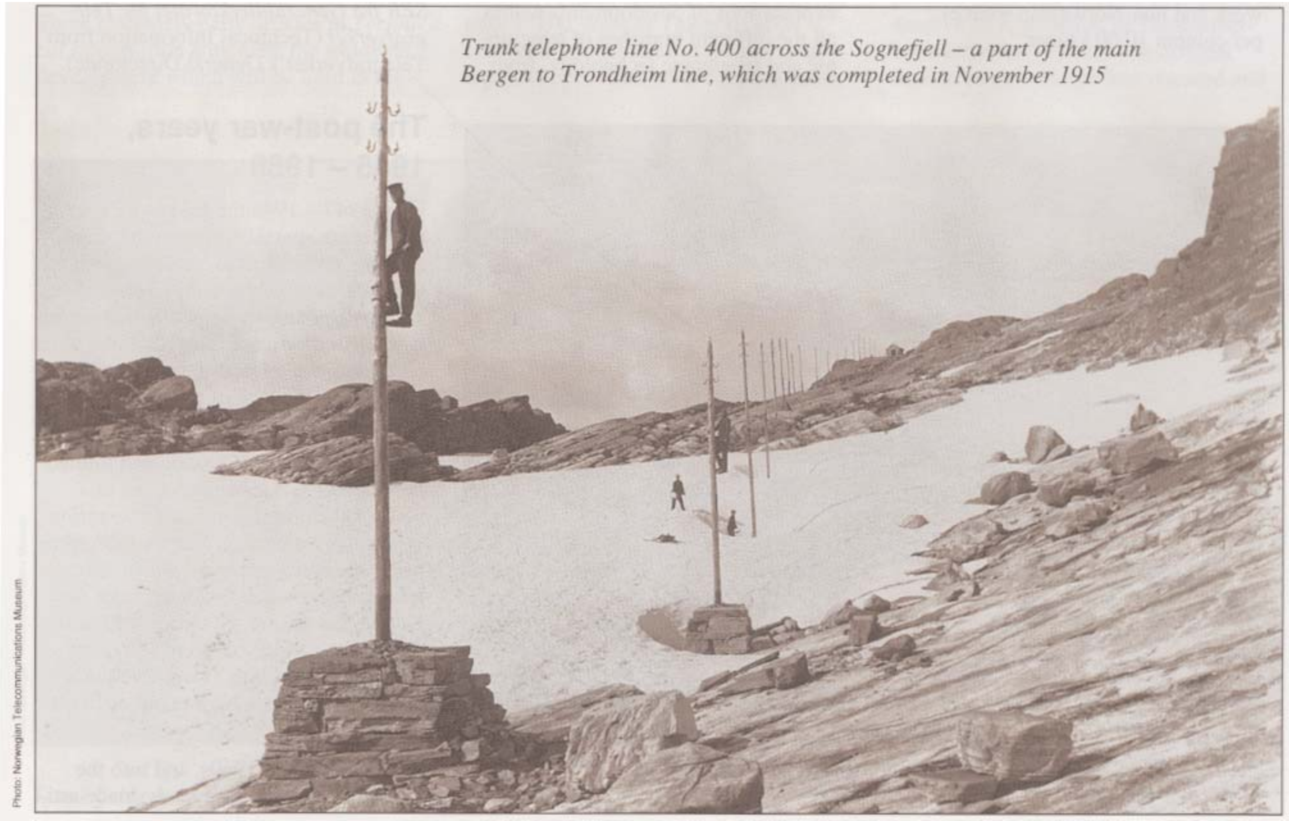
Interlude, 1916 – 1921

A new edition came out again in January 1916, with a tiny change to the title *TEKNISKE MEDDELELSER fra Telegrafstyret*.

It is at this point that ascribed articles first appear. As in the previous period, the contents are dominated by descriptions of new types of equipment and technical solutions in circuitry.

The idea was continued that *Technical Information* should appear once a month, and indeed this frequency was maintained until 1920. The three last numbers for this period were in fact marked up beforehand as belonging to 1921, but apparently were first printed in 1925 in order to complete an already commenced production run.

Again this journal takes a break, this time for five years.



Trunk telephone line No. 400 across the Sognefjell – a part of the main Bergen to Trondheim line, which was completed in November 1915

Photo: Norwegian Telecommunications Museum

Photo: Design Consult AS

This design set the trend for many years to come. A plaster cast of Jean Heiberg's bakelite telephone from 1934

Serious production, 1927 – 1945

Technical Information re-appeared with a new first issue in January 1927, under a more obvious editorial control, even though the editor is never named in the publication itself.

For the first time there is an incentive for contributing to the publication, in the form of payment:

> "For original work, which is assumed to be based on the author's own research, per column15.00 kroner
>
> For articles based on both original work and non-Norwegian sources, per column10.00 kroner
>
> For items translated from non-Norwegian sources, per column 7.50 kroner"

In his introduction to the first issue, the editor declared that "In those intervening years when *Technical Information* did not appear, there have been significant technical developments in both fixed and wireless telegraphy and telephony", and he goes on to say that:

> "In order to secure a common level of knowledge within our readership, we must lay the basic foundations first in order to build on them in future articles. So we must provide lengthier explanations of developments within all the different branches of telegraphic and telephonic technology, from the earliest beginnings to the later developments, as well as providing other items of interest."

Photo: Norwegian Telecommunications Museum

Broken telegraph pole – a picture of the ingenuity that had to be shown with the scarcety of resources in the reconstruction work after the Second World War, Finnmarksvidda in Northern Norway 1945

This objective is pursued over a considerable period of time, with long theoretical descriptions, occasionally stuffed with mathematics, which apparently some of the Telegraph Administration's own teaching staff enjoyed, so it seems, from an assumption that countrywide, a host of 'starving' engineers hunger for this schoolbook feast.

Technical Information appeared to have a certain unbalance in its coverage of the various specialist areas. At that time the three main areas of technology were: line/transmission technology, telephony, and radio. The first mentioned was predominant, but gradually, a range of subjects was added, to whit, the economics of technical development, and status information from the cooperative work in the international telecommunications organisations.

There was a little irregularity in production frequency, and occasionally, bundled issues, but the journal continued smoothly and sweetly through the thirties and even during the war. The latter had little apparent effect on the journal except that from 1943 to 1945 the name changed to *TEKNISKE MEDDELELSER fra Generaldirektoratet for Telegrafverket* (Technical Information from Telegrafverket's General Directorate).

The post-war years, 1945 – 1958

Until 1947 – 1948 the first years after the war are sparsely represented. Like all other professionals, those capable of technical authorship within Telegrafverket (the Norwegian Telecommunication Administration), are fully occupied with rebuilding the telecommunication networks after the lack of maintenance and destruction of the war years.

After many years of a static and limited portfolio of services came signs of a new awakening. *Technical Information* wrote about a new service 'A subscriber network for teleprinters', which was later to become telex (teleks under the Norwegian spelling system.) Automation of the service began in 1957, and telex developed into Telegrafverket's prestige service through many difficult years.

At the end of the 1940s, and into the 1950s, a new source of ready-made arti-

Photo: Norwegian Telecommunications Museum

Manual telex exchange, Oslo 1952

cles pop onto the scene. A steady stream of Telegrafverket's employees go abroad on study tours, especially in the US. Many articles from that time are based on information from these travels, that throw light on various subjects, lines of development and future possibilities for re-building and revitalising the telecommunication networks. There is hesitant mention of the potential for radio link networks in Norway, and possibly even television!

Three extremely early articles in *Technical Information* anticipate the technical development which will be used in the *application* of new technology. The first article in 1951 talks of the principles of PCM (Pulse Code Modulation), the second in 1953 discusses new switching principles in relation to telephone exchange ITT 8A, and the third in 1956 introduces the transistor as "a new aid in transmission technology".

In 1952 Telegrafverket's comprehensive achievements in supporting the VIth Olympic Winter Games in Oslo are described. A large centenary issue in 1955 called '100 years of Telegrafverket – characteristic features of technical development', provides an exhaustive status description of the state owned corporation's operations.

On 14–17 June 1956, the Third Norwegian Telephone Engineer's Conference is covered, with the articles heavily indebted to the lectures and follow-up discussions. The journal also finds the internal District Engineers' Meetings to be a good source of material.

The reconstruction of the Northern Norway telecommunications network is described, also the development in the use of overhead lines and buried cables. The provision of co-axial cable in Southern Norway is extensively documented in equal detail.

But not everything politically important for the technology sector is reflected directly in the columns of *Technical Information*, for example the important resolution in the Storting in 1953 concerning test television transmissions, and the later resolution in 1957 on permanent transmissions. There is no mention of these in any articles until much later.

Ramping up, 1959 – 1970

It was getting to be time for a name change. For many years, and especially after the war, the publication gradually adapted more and more so as to come to deserve its appellation as Telegrafverket's technical journal. The original limited subject matter was retained and

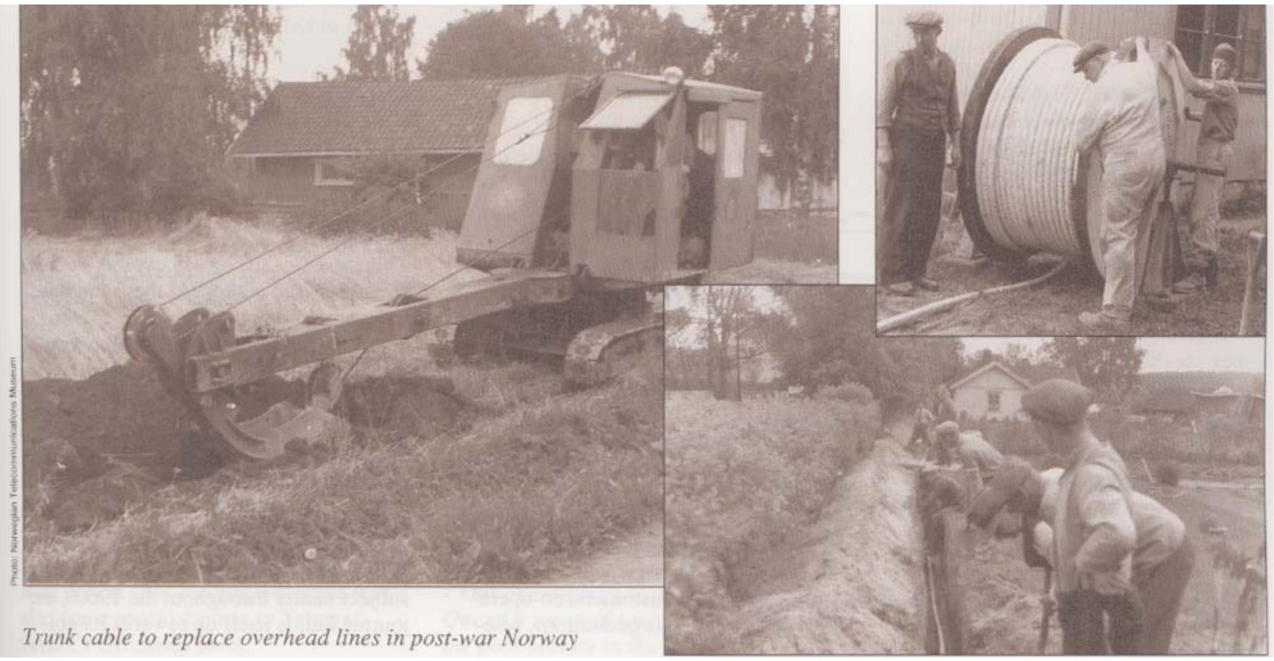


Photo: Norwegian Telecommunications Museum

Trunk cable to replace overhead lines in post-war Norway

The official opening of Norwegian television in the presence of King Olav and the Prime Minister, Einar Gerhardsen 20 August 1960

gradually extended. The need to document the intense development and changes which were involved in Telegrafverket's operations, were followed up.

From 1959 the journal is called *Telektronikk,* after a name competition. The editor pleaded

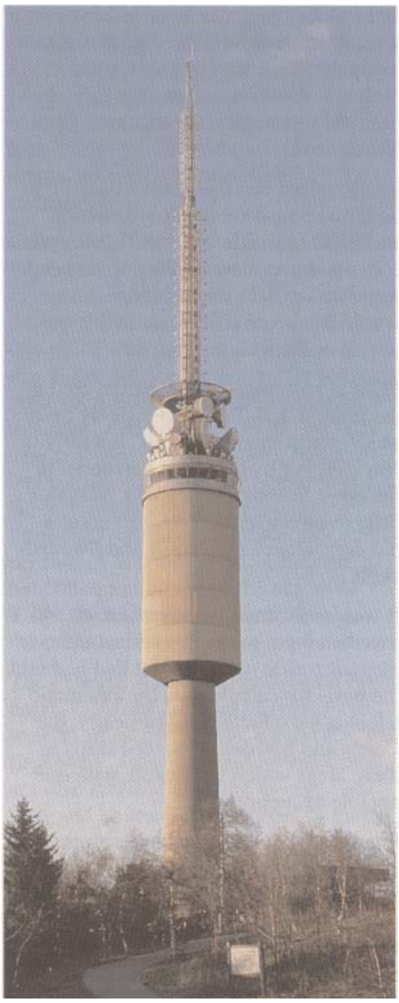
> "Yes, this is creating a new word in our vocabulary, but we hope and believe that it will introduce the new concepts contained in our journal.
>
> The new name begins with 'tele', which is beginning to gain ground here in Norway as a word that encompasses the idea of telecommunications.
>
> We also have the word 'elektronikk' (electronic) included, a word that most of our readership will be familiar with."

In the year previous to the change of name, the ambitious level of output of 12 editions per year was adjusted down to 4, but it was plainly intended that these should be comprehensive issues.

Throughout the 1960s and 1970s, the biennial Norwegian Telephone Engineers' Conferences – and also to some extent those of the annual Radio Engineers – continued to be important sources for articles for *Telektronikk.* These meetings were used to energise the whole electronic environment in Norway by cross-pollinating telecommunications, radio and television development, acoustics, etc. They provided a forum for suppliers and customers, co-operating infrastructure organisations, research institutes and technological universities. Also invited to the meetings were foreign experts from both research and industry and – to the Telephone Engineers' Conferences – representatives from the other Scandinavian PTTs.

This increases the areas covered, both by subject and author. At the same time lectures and reports from Telegrafverket's District Engineers' meetings continue to provide the 'bread and butter' articles. But there is also a continuous stream of articles either commissioned through editorial activities or spontaneously contributed, with perhaps a mixture of guilty enthusiasm or dutiful inspiration.

With a gradual increase in quality and subject matter throughout the 1960s, the



Tryvannstårnet (the Tryvann tower), radio link HUB and TV transmitting station. The tower was completed in 1961

Telegrafverket suffered a considerable drop in prestige when Forsvarets Fellessamband (Norwegian Joint Signals Administration), a defence agency, was set up in 1953, in reality as a competitor to the state monopoly. The main reason for its construction was Telegrafverket's unwillingness to consider the use of radio links as a means of connecting transmission networks, which had to be constructed extremely quickly using finance from NATO funds.

Even so, the laborious task of transforming a decrepit Telegrafverket to the modern Televerket is now in process, and carried through with grim determination. This indeed is reflected in *Telektronikk.* The presentation of a plan to develop a broadband network in 1959 is hailed as a saviour. Broadband radio link connections in the telephone network are a precondition both for speedy development of trunk dialling and for a core network for television and FM broadcasting.

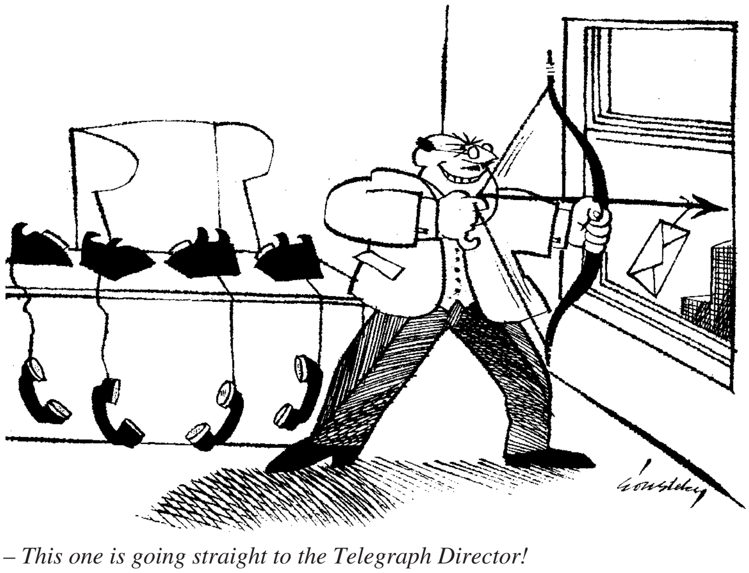
Telegrafverket comes in a little late, but proceeds with an enormous deployment of radio links. The radio technology part of the administration regains its representation again through articles in *Telektronikk.* The immense struggle to provide Norway with television is less documented, but all aspects of the construction of the radio link networks are written about enthusiastically in articles in 1962, 1963, and 1967. The 'radio men' were still allowed to be proud of their pioneering spirit.

There are also many articles on telephony automation. In one period (1962, 1964, 1965 and 1966) there are some reports on long distance call automation, which later become known as subscriber trunk dialling, and later just trunk dialling. Cross-bar exchanges for both local and trunk dialling are covered in an article in 1963. Two articles on traffic theory concerning Engset's formula are produced in 1966 and 1968, a subject which is later developed for a doctoral thesis.

New telephony systems were presented at two Telephone Engineers' Conferences. The first in 1962 (referred to in 1963) was what was later to become the ill-reputed ITT System 8B, and in 1968 a very important presentation by both of the main telephone exchange suppliers on their new SPC (Stored Program Controlled) telephony systems, namely Ericsson's AKE and ITT's 10C.

Early experiences with plastic insulated cables were discussed in 1962. Then comes a whole series of articles on subjects in the early stages of development that were later to become important to Telegrafverket: in 1962 on data transmission (early international collaboration), and in 1964 on satellite communications (Scandinavian co-operation had already begun in 1961), on the application of PCM, and in 1967 on 'The public land based mobile VHF radio telephone service', a rigorous exposition of the manual service which after automation was called more shortly mobile telephony. An journal builds itself up towards its golden age. However, there was one area which found no coverage in either *Technical Information* or *Telektronikk,* namely Telegrafverket's problems and suffering public reputation.

The waiting list for a telephone line had reached its first peak of around 78,000 in 1956, and decreased for a short while before it began to climb again towards a second peak. A flawed policy called 'the BMV plan' (Bring down the waiting list), led to an imbalance between the connection of subscribers and the development of the network. Subscribers experienced this as a lack of dialling tone, and unreliable grade of service. The main cause of the problems was too small governmental appropriations for a long period after the war, but there was also some speculation as to Telegrafverket's efficiency level.



– This one is going straight to the Telegraph Director!
One of a series of sarcastic newspaper drawings criticising the long waiting lists and the poor service in the 1970s.

Televerket's Research Institute (TF). The first building phase of TF's own premises near completion in September 1973

article in 1970 was an overview of plausible perspectives for telecommunication technology.

The earlier imbalance in the journal's coverage of various specialisms is finally evened out. But then new integrated technologies come in, blurring the traditional lines between disciplines. That is also one of the premises that were laid down in the total re-organisation of Telegrafverket in the early years of the 1970s, in which also Telegrafverket is re-named Televerket.

First peak period, 1971 – 1980

Televerket's reputation is perhaps not what it should be. There are complaints about the high prices for telephony, and the waiting-list only seems to increase, until it peaks in 1979. At that point it goes over the top, with 120,000 people waiting to be connected, who are for the most part from the domestic sector. Business customers have first priority. The worst problem for these latter subscribers is the poor grade of service in the trunk network, which is not at all satisfactory.

But the size of *Telektronikk*'s editions increases steadily, with a steadily increasing in-flow of articles. Soon there is no professional skill subject that is not covered in one or more articles. What are the reasons for the excess energy, all the enthusiasm within the company, despite the problems?

The re-organisation of Televerket basically revitalises the company and its internal processes, but while it continues it demands much resource and is a strain for many people. The re-organisation began at the top, when Televerket received a government appointed board from 1st January 1969. At the start of the 1970s, local units are reduced from around 150 to 27, and the number of regional organisations comes down to 6 from 12. The traditional specialist groups are torn apart in the head administration, which is strengthened in the areas of planning and finance.

A second factor in the new operational style is Televerket's Forskningsinstitutt, TF (Televerket's Research Institute), which was created at the end of 1967 after much heart-searching both in and outside the company. By establishing a research institute, Televerket proclaims its intention, both inside the company and to the world at large, that it intends to achieve more than just clearing the waiting-list for telephones and the problems with obsolete automation equipment.

With the benefit of hindsight as we consider the value of TF, it is useful to distinguish between the debate as to what TF qua research institute 'added to the agenda', and on the other hand what the new introduction of TF implied within the organisation. It is obvious that the influx of young, ambitious researchers necessarily had to lead to a new keenness in all the professional operations of the company, far beyond the technical areas alone. And this is indeed the case when later reviewed, for we can see that many, large projects are the result of close co-operation across many disciplines, and that many also appear at a Nordic level.

The first two prestige projects for TF in its initial phase are data transmission and PCM. Already in 1969 *Telektronikk* contained overview articles on both subjects. In the 1971 edition there is a whole stream of articles on PCM, and not all of them came from TF.

PCM is so comprehensively covered in the journal that it was extensively used as an educational handbook, not only in Norway, but by technical personnel in the other Scandinavian PTTs. Was this a little re-payment for Telegrafverket's old dependence on Danish and Swedish skills?

It was in no small measure Danish and Swedish input to the theoretical treatment of traffic management that had laid the basis for calculations for automatic telephone exchanges. Historically, the Danes and Swedes possessed skills in these areas which we lacked, with one notable exception.

Engset's formula, however, is an interesting example of original work of international standing that comes from Norway, and was published in 1918 in a German journal. (An English translation: 'The probability calculation to determine the number of switches in automatic telephone exchanges' was printed with commentary in *Telektronikk* No. 1.1992.) T. Engset was promoted to Director of the Telegraph Administration from 1930 to 1935, after a long career in the company, beginning as an eighteen-year-old in 1883.

More articles in *Telektronikk* bring Engset's formula forward into the limelight. Already mentioned are those in 1966 and 1968. The author, L.A. Joys, from the Telecommunications Administration in Bergen, in 1972 took a doctorate based on further development of Engset's work.

One of these articles is called 'A comparison of analytical methods and simulations for dimensioning telephone systems'. Simulation is one of the methods that TF had just started to use in the institute's third input area – switching and teletraffic. A practical application was needed shortly. In a collaborative effort between TF and technical units within the Teledirectorate and the regions, an offensive was launched to find a speedy solution to the problems of trunk telephone network. The problem solutions gave rise to many articles in *Telektronikk,* for example on grading, signalling investigations and traffic measurements, all with reference to 8B exchanges.

Europe is in need of traffic and O&M (Operations & Maintenance) measurements to serve as a basis for network design and planning. During the 1970s one of TF's largest research efforts is on this issue, putting the teletraffic research group at the international forefront at this time. Articles from 1978 and 1979 deal with operational control systems and traffic recording, the beginning of a subject that later was to become very important for telecommunications administrations and the object of international collaboration under the acronym TMN (Telecommunications Management Network).

The external Telephone and Radio Engineers' Conferences continue to be the sources of many articles, but the District Engineers' meetings, an internal forum, are discontinued after the re-organisation, and so one of the sources of material for articles disappears.

However, there is another possible reason for the increase in the flow of articles to the journal. With an administration under an appointed board, the new organisation of Televerket needs a higher degree of formalisation of work proposals, in comparison with previously. At the same time project work is becoming the norm in the whole company. With increasing need for documentation, planning and other proposal documentation may be used as a basis for articles for Telektronikk, even though it is not always easy to trace this origin.

Two examples from 1975 can be given: One article describes the whole range of the automated telephony systems in the Norwegian network (looking forward to 1982), and a second article describes the structure of the future digital telephone network. The articles are based on the preliminary work for the long-term plan of Televerket, which is completed by the turn of 1979.

The development of the digital telephone network is covered in a series of articles, beginning in 1972 with an article based on a Swedish input to the Radio Engineers' Conference in 1971, and thereafter covered as a subject in *Telektronikk* almost annually including articles on ISDN in 1977, 1978, and 1980. The ISDN concept (Integrated Services Digital Network) was formulated by a study group in CCITT in 1972 and was held up by many people over the years as the ultimate goal for all telecommunication network development.

Many stages in the process, from planning to experimental development and on to actual construction, are described in articles, in 1972, 1973, 1976, and 1978, about the Nordic partnership project NPDN (Nordic Public Data Network). Later, datex came to mean this line switched connection service. A competitive service, more internationally acceptable, for data transfer based on packet switching, was described in an article in 1980, some years before the service was launched as datapak.

The development of NMT (Nordic Mobile Telephone) is not as fully covered in *Telektronikk* as the data network project. As a Nordic partnership project it is even more successful and had a great following. In 1979 the first article on NMT appears, when the project is already at a late stage of development.

There is frequent coverage of various aspects of satellites and their application – already in 1967 an article on transmission delay in geostationary satellites, on the influence of the troposphere beyond the 10 GHz frequency (1975), on earth-station antennae (also 1975) and on to the first articles on NORSAT in 1976 and 1977. With the NORSAT system for transmission to oil-platforms in the North Sea and later on to Spitzbergen, Norway becomes the first country in Europe to use satellite communications for national purposes.

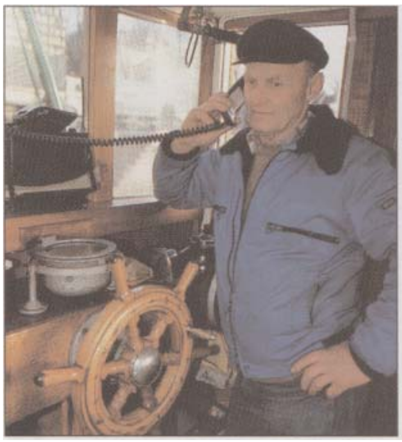


Photo: Rune Lislerud/SAMFOTO

An NMT mobile phone in practical use on board a small coasting vessel in the 1970s

Push buttons and multi frequency signalling was the key to a multitude of new telephone services. "Tastafon" was developed in the late seventies and introduced in 1981

In articles in 1978, 1979 and 1980 new telecommunications services are presented, telefax, teletex and teledata, respectively. With hindsight, telefax is a great success, teledata would never take off, but teletex was a regular fiasco.

Amongst all these modern items in *Telektronikk,* there was one article that harked back to a previous era, 'Accumulator batteries, a trusty work horse in Televerket's network'!

A review of part of the technical forum can be found in a broadly-based centenary issue in 1980 with the main leader being '100 years of Norwegian telephony'. In the introduction to the issue we find a paragraph "Televerket stands at a watershed before converting to new technology, with even more extensive changes than previously. The new telecommunications network is growing and displays the whole technical sector against the background of new and exciting possibilities and challenges." *Telektronikk* would not spend too much time on the past achievements.

The journal continues its reporting back on the plenary meetings of CCITT and CCIR, and exceptionally also from other international telecommunication conferences.

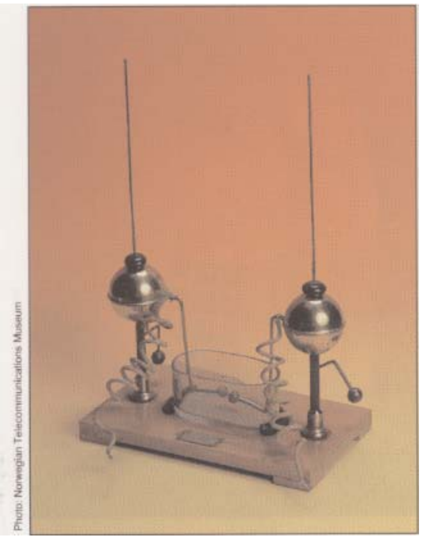


Photo: Norwegian Telecommunications Museum

Spark transmitter from 1905

In the consideration of all the material in *Telektronikk*'s volumes, obviously many subject areas cannot be recorded in this overview article, but thorough articles appeared on e.g. optical telecommunications in 1975 and 1979, the grounding of telecom installations in 1976, on satellite broadcasting and on SPC telephone exchanges in 1977, on aviation navigational systems and on new telephones in 1978, new developments in short- and medium wave transmission and also cable television in 1979, and on new cable types in 1980.

It looks as though a small attempt was made to thematise the issues. In 1972 there is an issue on the introduction of colour television, and in 1974 one with many articles on the theme of teletraffic systems.

Perspectives in teletechnology were covered in overview articles in 1972, 1974 and 1976. Right at the other end of the scale we find lectures with highly theoretical and in-depth discussions, not listed here. The overwhelming mass of articles deal with more or less practical subjects.

The time lag between important events in the technical sector and the journal's coverage of it becomes ever shorter, be that decisions, operations or actual achievements. Telektronikk thus becomes better matched as source material.

But the total impression of the material is still that of a kaleidoscope; the editing of individual issues and the succession of articles over a period, is accidental and mostly indicative of the date of manuscript production. Here is an area of work waiting in the wings – the systematic organisation of articles in *Telektronikk* by theme from early beginnings and through stages of development till final results.

Second peak period, 1981 – 1987

Televerket's long term plan was already completed and made public in 1980. The ground rules were laid for the debate which should provide a final clarification of the role of telecommunications in a modern society, and so remove all the financial problems for the company – so they thought.

Photo: Design Consult AS

Medium and short wave antennas at Kvitsøy transmitting station, on the air from June 1982

However, quite a few more premises were to emerge, in the form of a series of official reports initiated by the Ministry of Transport and Communications, and inspired by the wind of liberalisation that was blowing in the USA and had been taken up by the European Community, which in turn was pushing hard for European development. In Norway proposals from appointed public committees comes out with astonishing regularity in the official reports about Televerket every year over a five year period.

Thanks to the comprehensive preparation that has gone into Televerket's long-term plan, the company has ready made answers to all the questions that are raised in the public committees. For example, there is an article in *Telektronikk* in 1981 about utilisation and number capacity in automatic telephone exchanges, based on a statement to Teleutvalget (the Telecommunications Committee), which had been rather critical.

Telektronikk's annual output now gives the impression of an increasingly self-confident company. A long series of articles describes the good results in both well-established and new areas of expansion for the business. Perhaps the habit of contributing to the journal is becoming more established? The annual output is maintained at a stable size. Earlier, these occasions had been sporadic, but now we see more themed editions, i.e. where articles relating to a theme are edited together to provide the main contents. The editorial function becomes more visible.

There are two themed issues in 1981, the first containing three articles on satellite communications – one for each of INTELSAT, INMARSAT and NORSAT. Norway has a considerable share in INMARSAT because of its shipping interests. The second themed issue is on telex, on the occasion of a new completely digital exchange for the telex network, with coverage of the range of services and the standards for a prioritised business sector. In 1982 there is another themed issue on the expansion of the communications network in Northern Norway, including an article on the subscriber network in Longyearbyen, Spitzbergen.

A themed issue in 1986 contains a thorough run-through of most of the planning

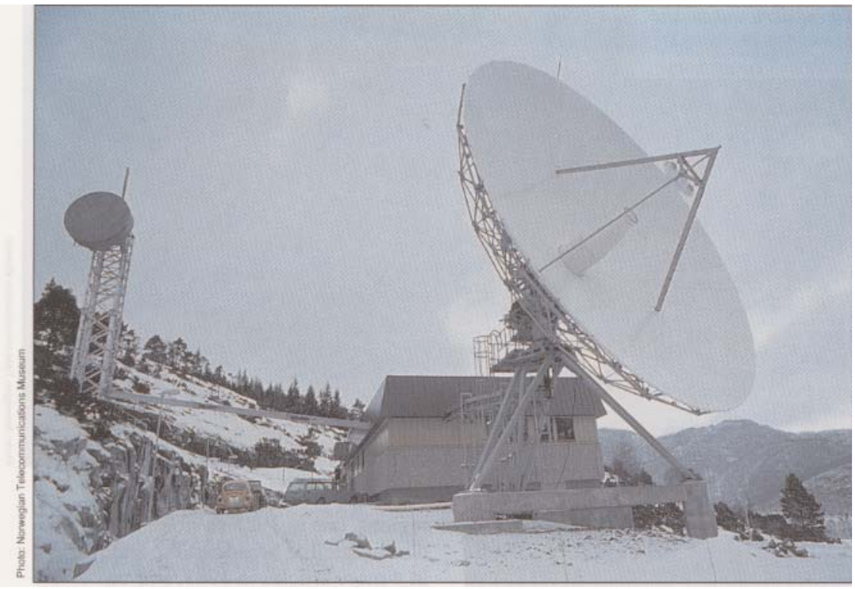


Photo: Norwegian Telecommunications Museum

First in the world as a fully automatic INMARSAT coastal earth station, Eik in Rogaland was operational from 1982

phases for technology and finance, and the strategies for the necessary training. The articles include fundamental policy statements at a high organisational level, right down to a run-through of practical aids to planning.

Back to the Telecommunications Committee, which sets its mark on *Telektronikk.* On its journeys in its deliberation phase, it visited France, where it was seduced by the idea of telematics. The committee touted the idea of building a broadband network, to be called a telematics network, with cable television installations as one of its elements. Televerket is given a grudging order to calculate the costs of this broadly outlined plan. In 1985 many articles give an account of the plan and its preparatory phases. In the same year, however, the Committee's concept of a telematics network is written off in the Parliamentary statement on the future organisation and activities of Televerket.

Otherwise, 1985 is a memorable year for Televerket, not because of any special celebrations, but the telephone waiting list is brought down to zero, and Norway becomes fully automated. The process of fully automating the telephone network had been delayed by political considerations, so the date of 1985 is a little bit of embarrassing window-dressing.

Televerket's reputation rapidly improves and soon the only criticism is of the tariffs, which are kept relatively high since the board has a policy of gradually becoming self-financing. After 1987 Televerket is de facto completely self-financing and independent of state resources.

Many articles in *Telektronikk* present new services: there are two in 1983, one on teleconferencing and one on a test service for videoconferencing. There are also articles on paging (1984 and 1986), two on mobile data services in 1986, and a series (1984–1986) on electronic funds transfer.

Televerket becomes involved in cable TV, which is reflected in two articles in 1983 and two in 1985. The latter two form part of the collection of articles in connection with the proposals for the telematics network.

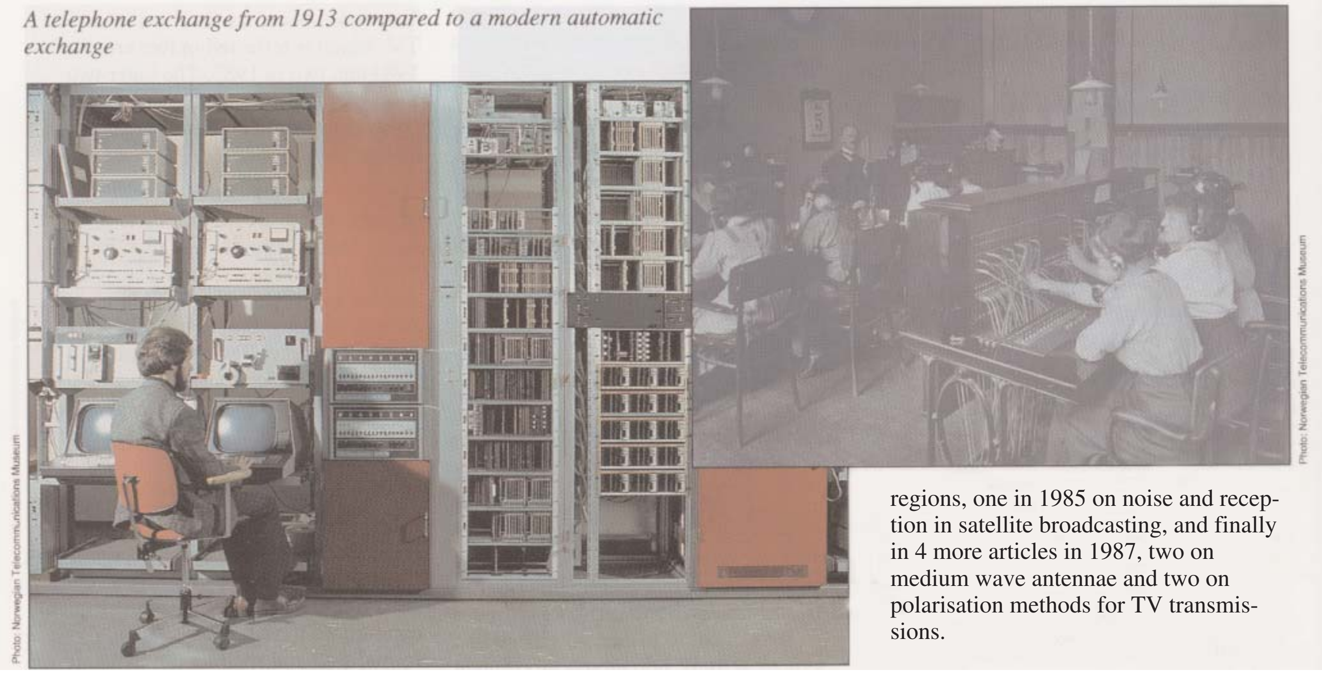
The Telecommunications Committee defined business communications as being open to competition in the future. So the journal also has articles on this subject in 1983, 1984, 1985, and 1986. As a prelude to competition Televerket starts Televerket Bedriftskommunikasjon, TBK (Norwegian Telecom Business Communications), as a project in 1984, and in addition to providing terminal equipment and business communications, TBK also takes over the company interests in cable TV.

It is indeed the Telecommunications Committee that mapped out Televerket's future development, as subsequent committees limit themselves mainly to formalising and spelling out in detail what the proposals meant.

It was the Telecommunications Committee's proposals that led to free competition in terminal equipment, which in turn led to repercussions on the organisation of the company, and also to Televerket shedding some regulatory functions. The Telecommunications Committee also led to political pressure being applied to the rate of replacement of automatic telephone equipment. Televerket had to put out an international request for bids for completely digital exchanges.

Photo: Norwegian Telecommunications Museum

Isfjord Radio, 78° N, on the island of Spitzbergen. Telecommunications to and from the island started in 1911. Experimental satellite communication started 1982



A telephone exchange from 1913 compared to a modern automatic exchange

Photo: Norwegian Telecommunications Museum

Photo: Norwegian Telecommunications Museum

In 1984 *Telektronikk* issues a comprehensive edition on ITT1240 – a completely digital telephony system, which won the bid competition. Twenty articles, covering almost 200 pages, describe a long series of features: the system architecture, management system and software, traffic and processor capacity, mechanical construction and installation, operations and maintenance, support systems, signalling, and so forth. However, it took two years before the first digital exchange was installed in Trondheim.

If there is a revolution in exchange operation, it is no less in the transmission area – it just takes longer. Digitalisation has rapidly increased in the local networks, but is held back for long distance. However, a necessary impetus comes with the deployment of a new generation of radio link connections. *Telektronikk* describes the new digital radio link transmission systems, which have a capacity of up to 1920 channels, in 1983, and in 1985 there are a further three articles on new digital radio link transmission networks.

Fibre cables was a new option at that time, but Televerket made careful preparations, with test installations in 1980 and 1982, and a year later laid the first ordinary fibre cable. Thereafter it gradually increased in penetration, although it was to take approximately 10 years before the backbone in the core network was completely converted. *Telektronikk* contained a series of articles on optical fibres, fibre technology and fibre cable: 3 in 1981, 2 in 1982, 1985, 1986, and 6 in a double themed issue in 1987.

The other theme in this latter issue is GSM, the new Pan-European digital mobile telephony system. The details of the international co-operation in this area are all given in the journal in 1985, followed by 4 articles in 1986, and then the double theme issue in 1987 consisting of 10 articles.

If we look at the subject of partnership work on the Nordic level, there are articles on Tele-X in 1985, on NORDSAT in the same year, and six articles on ISDN in 1987. On ISDN the Nordic PTTs have a certain degree of common policy and strategies. Tele-X is a much discussed project for launching a common Nordic satellite, and NORDSAT is an equally discussed proposal for common Nordic TV transmissions via satellite. It is perhaps the article in 1985 which provides the final insight and burial of NORDSAT, with its outstanding frankness in describing the project which had been driven forward by the Nordic Council.

Many articles cover radio transmission and broadcasting. Here are a few examples: two in 1981 on short wave transmissions in the Pacific Ocean, and on short wave broadcasting, one in 1982 on the Northern lights and radio wave transmission, two in 1984, both on the problems of FM broadcasting in coastal regions, one in 1985 on noise and reception in satellite broadcasting, and finally in 4 more articles in 1987, two on medium wave antennae and two on polarisation methods for TV transmissions.

Telektronikk continues to report the plenary committees in CCITT and CCIR, and gives detailed coverage of the Telephone Engineers' Conferences. Many more subjects are covered by the journal, but in this article it is possible only to comment on the most important development areas.

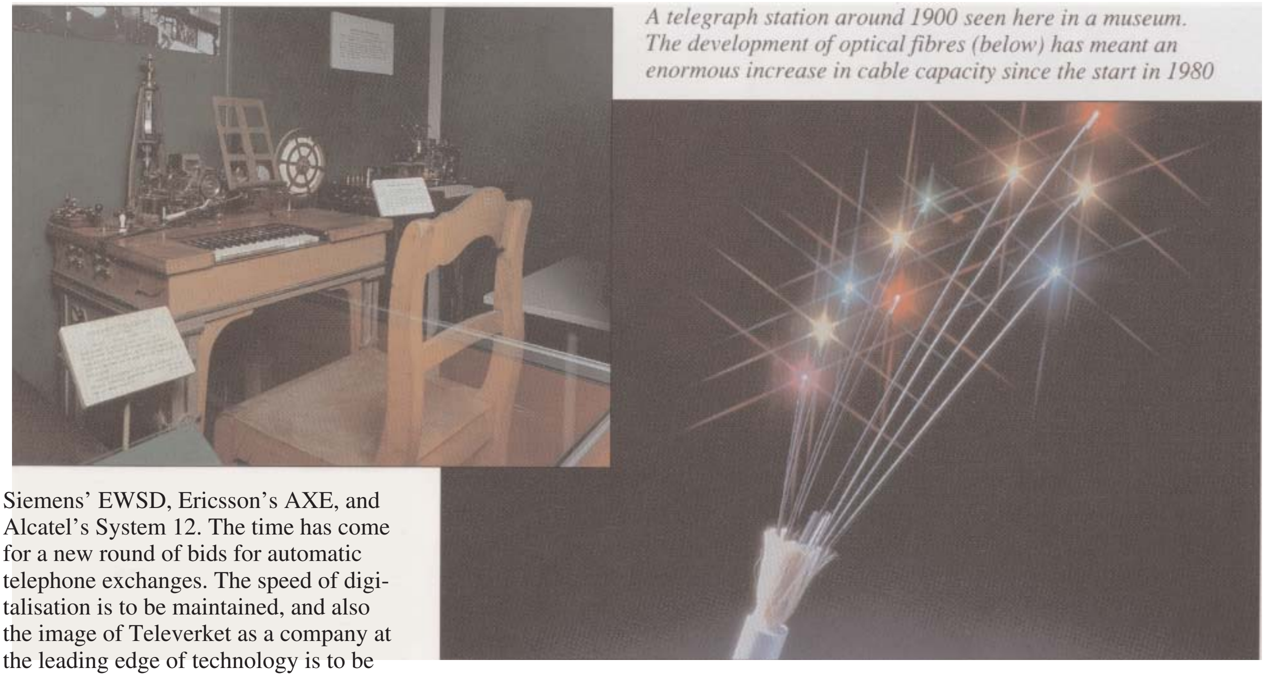
By the end of 1987 Televerket faces the organisational consequences of the intense public debates of the first half of the eighties.

Changing times, 1988 onwards

Competitive operations are split off from Televerket's core organisation after January 1st 1988, and are taken over by their wholly owned subsidiary, TBK A/S. At the same time regulatory responsibilities are transferred to a newly created body, Statens teleforvaltning, STF (The Norwegian Telecommunications Regulatory Authority). Now begins a long-lasting period of transformation to a new style of company.

If we look at the contents of *Telektronikk* we can see a tendency for a decrease in issue size. It is difficult to point to any one moment in time when this began but a 'slimmer' journal seems to coincide with the changes in Televerket's organisation and activities.

In 1989 three different digital telephone systems are reviewed in three articles:



A telegraph station around 1900 seen here in a museum. The development of optical fibres (below) has meant an enormous increase in cable capacity since the start in 1980

Siemens' EWSD, Ericsson's AXE, and Alcatel's System 12. The time has come for a new round of bids for automatic telephone exchanges. The speed of digitalisation is to be maintained, and also the image of Televerket as a company at the leading edge of technology is to be reinforced. The winner of this contract competition is Ericsson, and so the first AXE exchanges are installed in 1992 in Gjøvik, Trondheim and Bodø.

When it comes to the transmission sector, there are two articles on digital radio link communications in 1988, on optical transmission systems in 1988 and 1989, on optical components in 1989, and articles on fibre in the subscriber network in 1989 and 1991.

In 1991 *Telektronikk* publishes as a theme number, a series of articles on telecommunications in the Oslo Telephone District, covering a range of topics, from forecasting and planning to utilisation and optimisation in various sectors, the subscriber network, the local loop and the exchanges. There is also some discussion on the technical preparations for local broadcasting.

The moment of truth for ISDN is coming nearer. A test service is set up in 1990, with commercial operations planned for 1993. Both before and after the test service was established *Telektronikk* runs a series of articles covering a whole range of ISDN questions: six in 1988, four in 1989, six in 1990, and two in 1991. Many of the articles contain food for thought, and are a little hesitant. Similar questions are also raised in international circles, it is not a peculiarly Norwegian attitude: Is ISDN technology driven? What about applications?

Norway makes two bold commitments: the ISDN videophone, which is discussed in an article in 1989, and telemedicine which is covered in two articles the same year.

But the question of *utilisation* of the telecommunications network now occupies the journal far more than previously, with a change of emphasis in the contents of the articles from techniques and technology to applications. The old service ideas and principles are too restricting, there is a search for new ways to attack the problems, and to provide solutions. New acronyms continually pop up in articles – IN and UPT as well as the already well known VØT.

The first appearance of the idea of the Intelligent Network (IN) is in an article in 1988, which is followed by two more in 1989 and 1990. There are articles on each of Value-added Services (VØT) and Universal Personal Telephone (UPT) in 1990. Information security, a subject with a close involvement in service definitions, is covered in 1988. In 1989 there is a theoretical article on prime numbers, one of the basic elements for encryption.

If we consider abbreviations and acronyms, some of the essential new ones are: NICAM, DAB, RDS, ATM, and GPS. *Telektronikk* explains some of these terms with an article in 1989 on digital sound in television transmission (NICAM), articles in 1990 and 1991 on digital audio broadcasting (DAB), in 1990 on data transmission via FM transmitter stations (RDS), with two articles in 1991 on factors concerning transfer techniques via broadband channels (ATM), and finally in 1991 an article on terrestrial satellite navigation (GPS). Also to be mentioned are articles on Centrex in 1988, TMN in 1988 and 1990, and a new generation mobile telephone system (successor to GSM) in 1990. There may be a considerable range of subject matter, but the size of the journal is steadily decreasing.

One great item of interest is satellites and their communications applications, and Norway was indeed one of the pioneering nations in this area. In addition to the afore-mentioned article on particular satellite applications there is another on cable mapping via satellite in 1988, and in 1991 on communication satellites in new orbits, and finally, two overview articles. In 1990 there is a complete review, and in 1991 a review of mobile satellite communications.

As before, *Telektronikk* continues its coverage of references to NTIM (the Norwegian Telephone Engineers' Conferences) and overviews of the work of CCITT and CCIR and other international telecommunication fora.

But what is the reason for the stream of manuscripts into *Telektronikk*'s offices being reduced?

Another watershed is reached. A great period of revitalisation within Televerket is more or less complete, and the objectives which had been formulated at the

C-MAC, the new standard for transmitting sound and pictures with the use of multiplexed analogue components (MAC) was first used in Norway in 1984 for satellite transmission of television and radio to Spitzbergen

beginning of the 1980s are about to be reached, without being replaced by new goals which will awaken enthusiasm. Instead, a scenario of new threats is presented, which Televerket had to arm itself against.

The remedy of course is again to reorganise, and in addition to competition, there is now a new problem within the organisation: overmanning. The reorganisation would be a long, disturbing process which would affect every employee – directly or indirectly. To individuals this means less surplus energy, which is a prime condition to experts finding the time to write articles.

New deal in 1992

Televerket's technical journal began purely as a reporter on operations, cf. its original objectives, which are reviewed at the beginning of this article. Over the years the journal developed and gradually spread its coverage to include the whole national communication technical scene, and also made some inroads on the international environment. The Norwegian telecommunication environment was not of great size, chiefly because it was limited to the monopolistic Televerket, a few of Televerket's suppliers and NTH (Norwegian Institute of Technology). This situation continued well into the 1970s.

But the dividing lines between electronics, data- and information-technology are becoming blurred. The technical possibilities for telecommunications and information exchange open up such wide perspectives that they become areas of concern for both politicians and public.

So how should *Telektronikk* proceed? Should it compete in a large marketplace, should it change its profile, and how should it position itself in relation to scientific journals on the one hand and news-reporting journalism on the other hand, or should *Telektronikk* merely be an in-house circular?

Televerket is indeed in the midst of change, and as a result of this on-going process there is a new organisation from January 1st 1993. But that is only one stage in its development. Within two years it was hoped to organise it as a limited liability company. What would be the effect on *Telektronikk* in this move of Televerket from a controlled environment to market management?

These questions surrounding *Telektronikk* are thoroughly analysed and result in revised objectives:

1 *Telektronikk* shall be the leading Norwegian telecommunications journal.

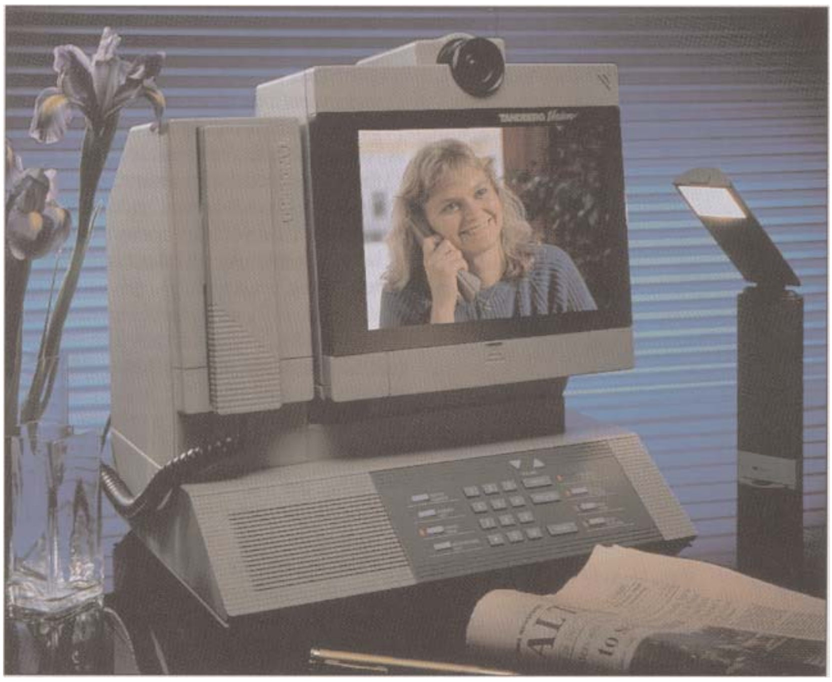
2 *Telektronikk* shall, through its choice of subjects and presentation format, contribute to synchronize professionals in the Norwegian telecommunications scene with reference to the development of telecommunications techniques.

The thorough development of an expanded range of subjects should now be the corner-stone of the journal's business ideal – its particular niche. Each issue will normally contain one main topic, which will allow the journal to position itself between the quickly outdated textbooks on new professional skills, and scientific papers and articles published in a range of journals, where the exposition can be more demanding. The articles in these themed issues will be written up in an educational style, so that there is no requirement for special knowledge amongst *Telektronikk*'s normal readership. Each issue will have a specialist editor with an international reputation. By limiting the contents to the selected topics, the journal can in great measure tap the reservoir of skills existing among the telecommunications professionals both in Televerket and the rest of the country – and partners internationally.

Telektronikk will also concentrate on international co-operation and standards work in telecommunications. As the former PTTs are preparing for competition in all their services, there is no less need for international co-operation, co-ordination and standardisation. The PTTs' transformation to competitors and operators of telecommunications services will certainly provide some constraints, but international partnerships and strategies, research and development, must all be in place, but at a pre-competition stage. The transfer of regulatory functions from the PTTs has led to new players entering the arena of international co-operation. Of the international telecom organisations, some of them are reorganised, some have disappeared and a few new ones have entered the arena. No longer is it an easy task to be aware of the international developments and to stay informed. Amongst the employees being assigned to different international tasks, *Telektronikk* has appointed a group of rapporteurs. Their reports are processed and systematized before appearing in the journal.

The journal has had a face-lift. Graphic design experts have created a new front cover of outstanding quality, which is a basic pattern of traversing lines with sophisticated use of colours, and will ensure that the journal has a well-defined image in the output of journals. The cover's 'graphical notation' is used as the basis for the designer's variations on a theme inspired by the main topic in each of the journal's issues. The layout of the text and illustrations has also gone up in quality.

There is now more continuity in the editorial style, with four sections with abbreviated titles, viz.: Feature, Special, Status, and Kaleidoscope. Already discussed are the Feature and Status (international activities) above. The Special section is available for articles of variable content at different skill levels, often spontaneously written by the contributors, i.e. like the articles that were contained in the earlier editions of *Telektronikk.* Kaleidoscope is regarded as a section for series or topics which do not easily fit into the other sections, e.g. articles on historical aspects, or commenting on pioneering scientific work in telecommunications and information theory (Maxwell, Erlang, Shannon a.o.).



The videophone developed through a joint co-operation project between TF and Tandberg Vision was completed in 1991

The output continues to be four issues per year. The first in the new style, No. 1.92, contains tailor-made articles in all the four sections, and has optical networks as its theme. The Special section contains articles on switching theory as applied to network structure, on the conversion to continuous or centralized charging, and a very useful article on the precise use of specialist terminology in high-speed communications. The Status section covers two new Pan-European partnership organisations, EURESCOM and ETSI. The Kaleidoscope section is occupied by a discussion of Engset's formula. Nos. 2.92 and 4.92 contain original work on their chosen topics, namely Intelligent Networks, and Satellite Communications, and have been produced in English. No. 3.92 is a special edition containing a full report of the proceedings at NTIM 1992, the 21st Norwegian Telephone Engineers Conference, which also marks the institution's 40th anniversary.

The 1993 series opens with an exclusive issue on Telemedicine. This application of telecommunications, based in Tromsø, was an ambitious TF project which aroused considerable international interest. No. 2/3.1993 is a double issue on Information Systems, and is limited to a consideration of telecommunications systems software. Several articles refer to the international recommendations and extensive coverage is given to TMN. Cyberspace is the rather speculative title for the last of the 1993 issues and concerns itself with multimedia. The articles cover the increasingly global networks of PCs and discuss possibilities and development trends, and also cover virtual reality, which inhabits the borders of telecommunications. In that issue the Status section discusses the logistics of reporting back from international standards fora, with an overview of the division into study areas, and the titles of all the rapporteurs in the specialist subjects. All the 1993 issues are produced in English.

Telektronikk's 90th year begins with a large edition, No. 1.94, concentrating on the topic of Forecasting, and is almost a textbook on the subject with 169 pages, containing practical examples on the application of prognosis theory to telecommunications. The feature section is in Norwegian, but the rest in English. No. 2.94 is about Broadcasting, and most of the articles talk about the changeover to digital technology in the different areas of technical radio and TV production.

No. 3.94 features a rarely covered topic, being a little different from the other issues, namely Arctic Telecommunications, and covers the experience gained during the construction and operation of telecommunication systems in Spitzbergen and Greenland, and also the Russo-Norwegian joint project for building a modern telecommunications system in North-West Russia. The last issue in its celebration year, No. 4.94 is on Electromagnetic Compatibility, which is covered from the points of view of both theory and practice, as seen from construction and production to operational responsibility.

All the editions in this 90th anniversary year contain reports and orientation from international standards fora and joint research projects, and the Status sections consist respectively of 17, 19, 35, and 11 pages. The first and last issues contain Special sections – on broadband services in No. 1.94 and in No. 4.94 one article on travelling wave antennae and two articles on IT, one covering information retrieval and a second on database management systems. One last article deals with a mathematical tool for network planning and routing.

What about the production team?

Being editor of either *Technical Information from the Telegraph Administration*, or *Telektronikk* (after 1959) has always been a sideline in addition to a normal position within the company. In chronological order the editors have been:

1904 – 1909	Hermod Petersen
1916 – 1921	Johannes Storstrøm
1927 – 1942	Sverre Rynning Tønnesen
1942 – 1957	Julius Ringstad
1957 – 1978	Nils Taranger
1978 – 1991	Bjørn Sandnes
1991 –	Ola Espvik

Thumbnail sketches of each editor will be found beneath their pictures, but it is interesting to note that they all have a common interest in, and dealings with, education and training.

Hermod Petersen, the first editor, only produced the journal for 6 years, as explained earlier. It is likely that this had some connection with his work as a 'radio man' during the hectic pioneering years, when Telegrafverket embarked on radio technology. More information on that can be found in *Telegrafverkets historie* (The history of Telegrafverket, 1855 – 1955) by Thorolf Rafto.

Most of the other editors were recruited from the telegraph or transmission areas of the business, as a consequence of the historical development of Telegrafverket's main businesses, the telegraph service and the long distance telephone service. In the beginning the administration was indifferent as to the local loop. After the private telephone companies were taken over, the administration proceeded to split the telegraph and long distance telephone from the local loop telephone service in many places. In Oslo, for example, the two local administrations were only brought together as late as 1958.

In addition to being so busy, Hermod Petersen as a 'radio man' would have had problems in collecting the 'correct' material for the journal. This is also mentioned in the *History of Telegrafverke*t on page 398:

> "In 1904 the administration started a technical journal, *Technical Information*. This should publish the progress and inventions in technical fields. Several issues contained items of interest to all and were well received, but there was a criticism around that the personnel had received 'stones and not bread', especially considering the lack of textbooks on telegraphy and telephony."

The second editor, Johannes Storstrøm, had no great luck in his efforts, even though he came from the 'right background', and he also has a reference in the company history:

> "What the journal contained was indeed excellent, but *T.M.* did not fulfil its promises. The new plan for 1916 to produce an issue every month, could not be achieved, and the material was not as comprehensive as could be wished."

When the journal disappeared for another six years, in 1921, it is not hard to see the reason why if one looks at the date. 1921 was the beginning of a serious economic crisis in Norway, and opened an era of restrictions, strikes, lack of exports, and bank failures. Telegrafverket itself, additionally, had suffered a series of swingeing cuts in all areas.

When Sverre Rynning Tønnesen became the third editor and took responsibility for the journal, he got it going again in 1927 and instilled some sort of order. With his yen for the theoretical and interest in education, and his later 'international inclination', he must also take some responsibility for having begun the long development of the journal in taking it to a higher level and broadening its subject matter. Many thought that this was at a cost to the 'bread and meat' for Telegrafverket's technicians, but it led to the journal widening its distribution far beyond the company's limits.

The drawing office in the Technical Department helped with the illustrations, a job that became an editor's secretary's sphere. Here is a list of them in chronological order:

1934 – 1963	Peder Stenseth Kleppestø
1964 – 1975	Olav Jensen Aares
1975 – 1991	John Sigvald Johnsen
1992 –	Gunhild Luke

The first three had all been leaders of the drawing office in the Technical Department, but when a new editor was employed within Televerket's Research Institute (TF), naturally the TF drawing office took over the illustration work. The present editor's secretary also functions in the Information section within TF.

From 1957 an editorial committee was created to support the editor, and its members have been:

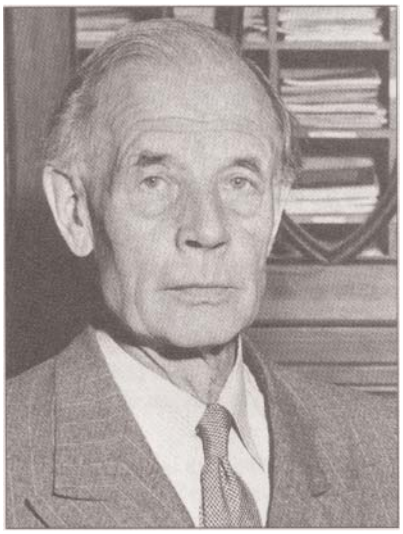
1957 – 1966	Julius Ringstad Nicolai Søberg
1967 – 1969	Nicolai Søberg Karsten Lagset
1969 – 1973	Per Mortensen Karsten Lagset
1973 – 1986	Per Mortensen Dr. Nic Knudtzon
1986 – 1990	Dr. Nic Knudtzon
1990 – 1991	Dr. Nic Knudtzon Ole Petter Håkonsen
1992 –	Ole Petter Håkonsen Karl Klingsheim Bjørn Løken

The directors of the Technical Department sat on the editorial committee until 1973. Then the Director of Research was counted as being on the 'technical side' in the management, following a reorganisation of the head administration.

The nomenclature 'editorial committee' was somewhat misleading as to the body's function, and from 1992 was corrected to a more fitting 'editorial board', and at the same time extended to include a chief of research.

Editor 1904 – 1909

Hermod Petersen, with education sponsored by the company, graduated from Karlsruhe Technical University in 1900, was Headmaster at The Telegraph Administration's Training Institute in Christiania 1900 – 1913, an Engineer in the Telegraph Headquarters Administration from 1913, Senior Engineer and Head of the Radio Department (later the Radio Office) from 1920, Chief Engineer and in charge of the Technical Department from 1931, and Director of the Telegraph Administration 1935 – 1938. Hermod Petersen was a radio technology expert of international repute. He worked on the tests which allowed Telegrafverket to open a wireless connection between Røst and Sørvågen in 1906, by which Norway became the second country in the world to have radio connections as a permanent feature in its telecommunications network. In 1911 he spent the winter as installation leader when the world's first Arctic radio station opened for business in Spitzbergen. In 1922 he was again busy with pioneering work building Norway's first broadcasting transmitter.

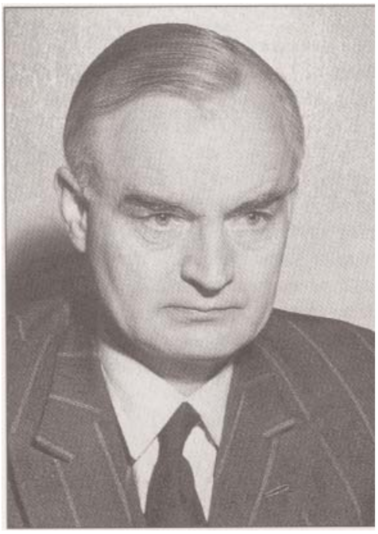


Editor 1916 – 1921

Johannes Storstrøm, with education sponsored by the company, graduated from Hannover Technical University in 1905, was an Engineer in the Telegraph Headquarters Administration from 1906, promoted to Senior Engineer and Head of the Construction Office, later Chief Engineer and Head of the Technical Department (later the Line Technology Department) 1935 – 1951. From circa 1930 his name was closely connected with the introduction of new 'electric' teleprinters as part of the modernisation of the telegram service. Johannes Storstrøm involved himself for many years with the education at the Telegraph College and wrote many textbooks on technical skills.

Editor 1927 – 1942

Sverre Rynning Tønnesen, with education sponsored by the company, graduated from NTH (Norwegian Technical University) in 1922, was an Operations Engineer at the Telegraph and Long Distance Telephone Station in Christiania from 1923, Engineer in the Telegraph Headquarters Administration from 1926, Senior Engineer from 1935, and Head of the Construction Office from 1939. He involved himself with education at the Telegraph College, and amongst other activities, wrote a textbook on line transmission. Rynning Tønnesen also actively participated in international telecommunications collaboration. During the war he was called to London, in 1942, where the Norwegian government appointed him Director of the Telegraph Administration to secure Telegrafverket's interests outside German-occupied territory. After the war he continued in this position until he retired in 1962.



Editor 1942 – 1957

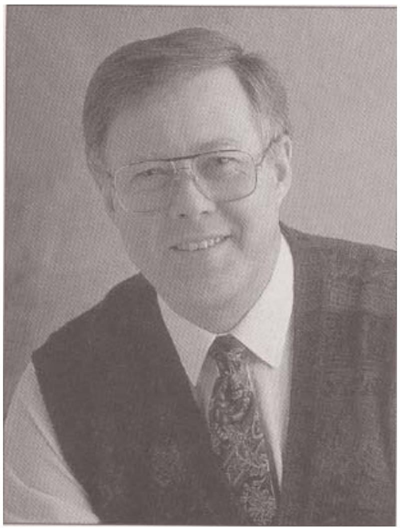
Julius Ringstad, with education sponsored by the company, graduated from NTH in 1924, worked for a short time at Møre district office in Molde, Operations Engineer at the Oslo Telegraph and Long Distance Telephone Station from 1926, Engineer in the Telegraph Headquarters Administration from 1939, Senior Engineer and Head of the Construction Office (later the Transmission Office) from 1945, Chief Engineer in 1956 and Director of the Line Technology Department from 1959, until he retired in 1966. Julius Ringstad was for many years a part-time teacher at the Telegraph College.

Editor 1957 – 1978

Nils Taranger, with education sponsored by the company, graduated from Dresden Technical University in 1943, served at the Drammen District Office, Engineer in the Telegraph Headquarters Administration from 1946, Senior Engineer from 1961, Chief Engineer and Head of the Transmission Office from 1967, Section Head of the Line Section from 1962 until he retired from Televerket in 1981. After the war Nils Taranger was in charge of the construction and automation of the telex network. For many years he was involved with education at the Telegraph College.

Editor 1978 – 1991

Bjørn Sandnes, with education sponsored by the company, graduated from NTH in 1963, was an Engineer in the Telegraph Headquarters Administration at the Transmission Office from 1963, for two years an Instructor at the Telecom College. He became Senior Engineer and Head of the Line Transmission Office from 1972, Chief Engineer and Head of the Telex Group from 1977, Leader of the Telephony Unit, later the Trunk Network Division from 1983. He was on leave for many activities abroad, the longest duration being 1977 – 1980. From 1994 he became Director of Corporate Strategy & International Affairs.



Editor from 1991

Ola Espvik, graduated in physics from NTH in 1968, is a research scientist at Televerket's Research Institute (TF) since 1970. He has been project leader of a series of research projects in the area of traffic and reliability dimensioning as well as operational control of the telecommunications network, and has been an active participator in international research cooperation. On leave from TF he was engaged by Televerket's Personnel Department from 1985 – 1988, to develop a training programme for network planning, and was Director for the Joint Technical College Centre at Kjeller in 1988 – 1989. Since 1985 he has also been a Lecturer and Study Advisor in telecommunications and computer science at UNIK – the University of Oslo's graduate Study Centre at Kjeller.

Has the journal achieved its aims?

In this short overview article there is not space for more than a cursory analysis as a finale.

The objectives for the journal were spelled out three times: in 1904, when it originated, in 1927 with its re-birth, and in 1992 when it was revised after the last change of editors. The formulations are referred to in the examination of the annual publications, but the essence of it can be gleaned from the references collected here for the sake of continuity:

> 1904:
> "...It is intended that this will impart knowledge to all personnel of the State Telegraphs, and will, in short, supply the more important news in the areas of telegraphic and telephonic technology."
>
> 1927:
> "In order to secure a common level of knowledge within our readership, we must lay the basic foundations first in order to build on them in future articles. So we must provide lengthier explanations of developments within all the different branches of telegraphic and telephonic technology, from the earliest beginnings to the later developments, as well as providing other items of interest."
>
> 1992:
> 1 *Telektronikk* shall be the leading Norwegian telecommunications journal.
> 2 *Telektronikk* shall through its choice of subjects and presentation format contribute to synchronize professionals in the Norwegian telecommunications scene with reference to the development of telecommunications techniques.

The objectives from 1904 and 1927 were published, but that of 1992 was set out in an internal memo approved by the editorial board.

As mentioned earlier, when discussing the editors, the evaluation seems obvious enough for the first two periods of the journal's existence, it did not perform and so disappeared. The choice of material was too biased, according to the company history by Thorolf Rafto.

After the new beginning in 1927, it seemed as though it had got organised, with an unbroken production through to the present day, but its sights were set too high as regards the number of issues per year. But having got the contents right, it was sensible to adjust the production down from 12 to 4 each year.

Why did people write? If we look at the first series of articles for Telegrafverket's technicians, and especially after 1927, it was with an almost idealistic desire to inform and improve their technical ability. The second impetus was the need for the local administrators and technical middle management to have some technical knowledge to be able to understand the tremendous developments, with the introduction of multi-channel systems and the change from overhead lines to cables. The description of multi-channel systems in *Technical Information* was often prescribed texts in the Telegraph Administration's educational courses.

But the gradual change in the journal began, a change which took a long time and was especially apparent with the mobilisation of the technological environments in post-war Norway. The increasing number of engineers in Telegrafverket with higher education are a target group for the journal, and with the raised level of content, the journal becomes of interest to environments other than the company. It is also distributed abroad, where it gains a good reputation.

Other countries' journals have certainly also been of influence, and over these years *Telektronikk* has had much in common with its sister journals in the Swedish and Danish PTTs. Perhaps it can also be seen to have some relationship to corresponding journals in other countries as well, namely the Netherlands, Switzerland and Australia.

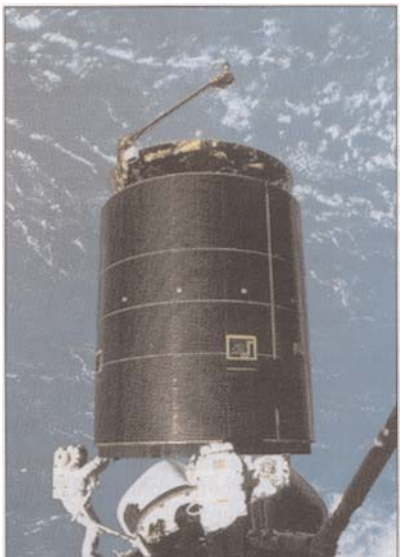
Which language should be used in *Telektronikk* – Norwegian or English? The gradual changeover to full production in English has come about amongst the other changes. This changeover seems to be an inevitable consequence of the journal's development, English being the

Some of Telektronikk's front covers since 1992

world language for technical matters. Reaching a wider audience with articles in English is of course an important incentive for writers.

Has the journal achieved its objective? The interest demonstrated both in Norway and abroad is very encouraging. Educational institutions ask for sets of copies to supplement other training material. The same goes for manufacturers working in the area covered by related themes. And serious professionals report back their use of the journal.

It is still too early to read the signs as to how the journal will succeed in the future. Producing themed issues is in accordance with its original ideal: to be educational. Now, however, that is achieved by entering a more all-embracing and higher level than expected originally, so the objectives of 1992 can be seen to have drawn level with *Telektronikk*'s development.



The technology has come a long way.
A telecommunication satellite is repaired in space June 1992