Telektronikk 1.93

Telemedicine

Dette eksemplaret må ikke fjernes!

Innhold

TEMA

SPESIAL

Telektronikk
Televerkets tekniske tidsskrift

89. årgang Nr. 1 - 1993
ISSN 0085-7130

Redaktør:
Ola Espvik
Tlf. + 47 63 80 98 83

Redaksjonssekretariat:
Gunhild Luke
Tlf. + 47 63 80 91 52

Adresse:
Telektronikk
Televerkets Forskningsinstitutt
Postboks 83
N-2007 Kjeller

Redaksjonsråd:
Teknisk direktør Ole Petter Håkonsen
Forskningsdirektør Karl Klingsheim
Strategidirektør Bjørn Løken

Utforming:
Gunhild Luke, Britt Kjus, Åse Aardal
Televerkets Forskningsinstitutt

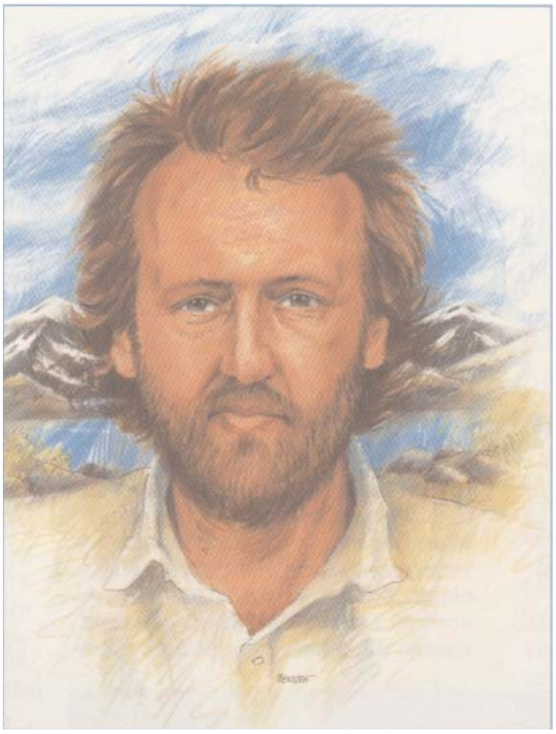
Tele RESEARCH

Telektronikk utkommer med fire nummer pr år.

Guest editorial

BY BIRGER J NYMO

621.39:61



Telecommunication services have traditionally been developed with a view to a mass market consisting of "anonymous" users. The stress has been put on technical solutions which promote optimum services in a perspective of equality.

The perspective has changed, partly because of the technical development, in the way that the particular needs of certain user groups are taken into consideration. Thus, the users of telecommunications no longer emerge as an anonymous mass, but in market relations they appear as household subscribers, company subscribers, small companies, large companies, etc. The services are adapted to the common characteristics of aggregated groups.

For the telecommunications of the future it is not enough to take into consideration the aggregated common characteristics. Individuals, groups, companies, etc. are inter-linked in social networks for co-operation, job sharing, community, etc. The composition and tasks of these networks are just as significant characteristics of the individual user as they are common characteristics of the group he or she belongs to. In order to offer optimum telecommunication solutions, this type of information about the subscriber must be obtained.

I would not be strictly accurate if I was to say that it was this recognition which formed the basis for starting a major research project on telemedicine under the management of Norwegian Telecom Research in the autumn of 1988. The basis was rather the fact that the Norwegian public health service was a major subscriber to telecommunications, but not a very advanced one. By realising individual applications the project has brought forward examples of a more advanced use of telecommunications. Along the course of the project the knowledge of the public health service of the present day and its future challenges have shown us the importance of flexible communication solutions on the basis of the tasks and structure of the public health sector.

The public health service has a lot to profit from efficient use of telecommunications. The greatest challenge the health service is faced with, is the growing gap between the expectations of the population, and the resources that society can set aside for health purposes. Increasing the level of efficiency and co-operation is therefore the most probable response to the challenge. Telecommunications may contribute to a more effective utilisation of resources through tying the resources of the health sector's resources together in a large number of telemedical services.

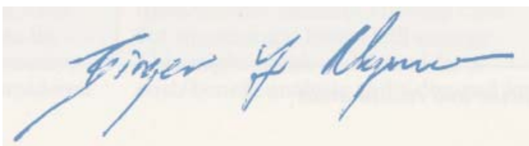
In order for the health service to profit from telemedicine various parties must set to work on a number of challenges.

On the political level, telemedicine must be integrated into the strategy of health care policies. Individual applications may relieve acute situations like the problems arising because of a geographically unequal distribution of medical specialists. However, the greatest profit will be gained when telemedicine forms an integrated part of the composition of the public health service. A significant piece in this context is the work that the Ministry of Health and Social Affairs has initiated with a view to laying the channels and regulations for the exchange of information in the health sector by establishing standards.

For the health services and the professions which have their daily work within it, it is a challenge to think in the way of new co-operative relations which demand different ways of organising work and thereby new work routines and roles. As sub-specialising within the medical professions continues, it will be necessary to think in the way of extensive co-operation beyond the administrative limitations which regulate the health services today.

The public health service is highly technological when it comes to medical equipment. New equipment offers the possibility of new methods of examination and forms of treatment. The challenge for the producers of this equipment is to ensure that the information produced by the equipment may be presented not only locally, but also be transmitted to experts at other locations in the health service for interpretation. Exchangeable data from medical information systems are also crucial in order to achieve efficiency through co-operation beyond the limitations of institutions.

The main requirements made to telecommunications by telemedicine are connected to flexibility and security. The health service constitutes an infrastructure which captures the whole of the population in a system from primary health service on the municipal level, to specialised hospitals with a national responsibility. The challenge is to match this medical professional frame with a flexible telecommunications structure which in a cost-effective manner enables the medical professional resources to be utilised wherever the patient is located. Security in this connection affects the protection of privacy as well as general quality assurance of medical services. On the one hand sensitive medical information must not go astray, and on the other hand the flow of information to the right person at the right time must be safeguarded without loss of quality.



Telemedicine

BY BIRGER J NYMO

621.39:61

Introduction

The concept of Telemedicine is not clearly defined. In the literature there are several definitions, which largely have been modified in step with access to teletechnology. In 1975 Bird stated this definition: "Telemedicine is the practice of medicine without the usual physician-patient physical confrontation via an interactive audio-video communication system". In the early eighties Conrath et al (1983) gave a far more general definition: "Telemedicine is the use of telecommunication technology to assist in the delivery of health care". Gradually, as the division between telecommunication and the management of information became more veiled, even this definition is too narrow, even if it takes into account other relations in the health care than only the meeting between physician and patient. What is called medical computer science is an important part of the technological angle of telemedicine. Basically, medical computer science reflects the application of data technology where storage, systematising, management and filing of information internally in medical institutions are superior factors. Such applications are a condition for telebased collection and reception of information whether it is for administrative or purely medical purposes.

In many contexts the concept of telemedicine is reserved for applications where the subject is to render health services based on application of telecommunication. Classical in that sense are forms of remote consultations and remote diagnoses within various medical specialities. In some cases transmission of knowledge in the form of distance education and remote instructions is included in the concept of telemedicine.

For our purpose we have found the following definition useful: "The investigation, monitoring and management of patients and the education of patients and staff using systems which allow ready access to expert advice and patient information no matter where the patient or relevant information is located" (AIM 1990).

The concept of telemedicine leads to many protracted academic discussions. Most important for practical work with telemedicine is to associate with the three main dimensions:

- telecommunication
- medical computer science
- health services.

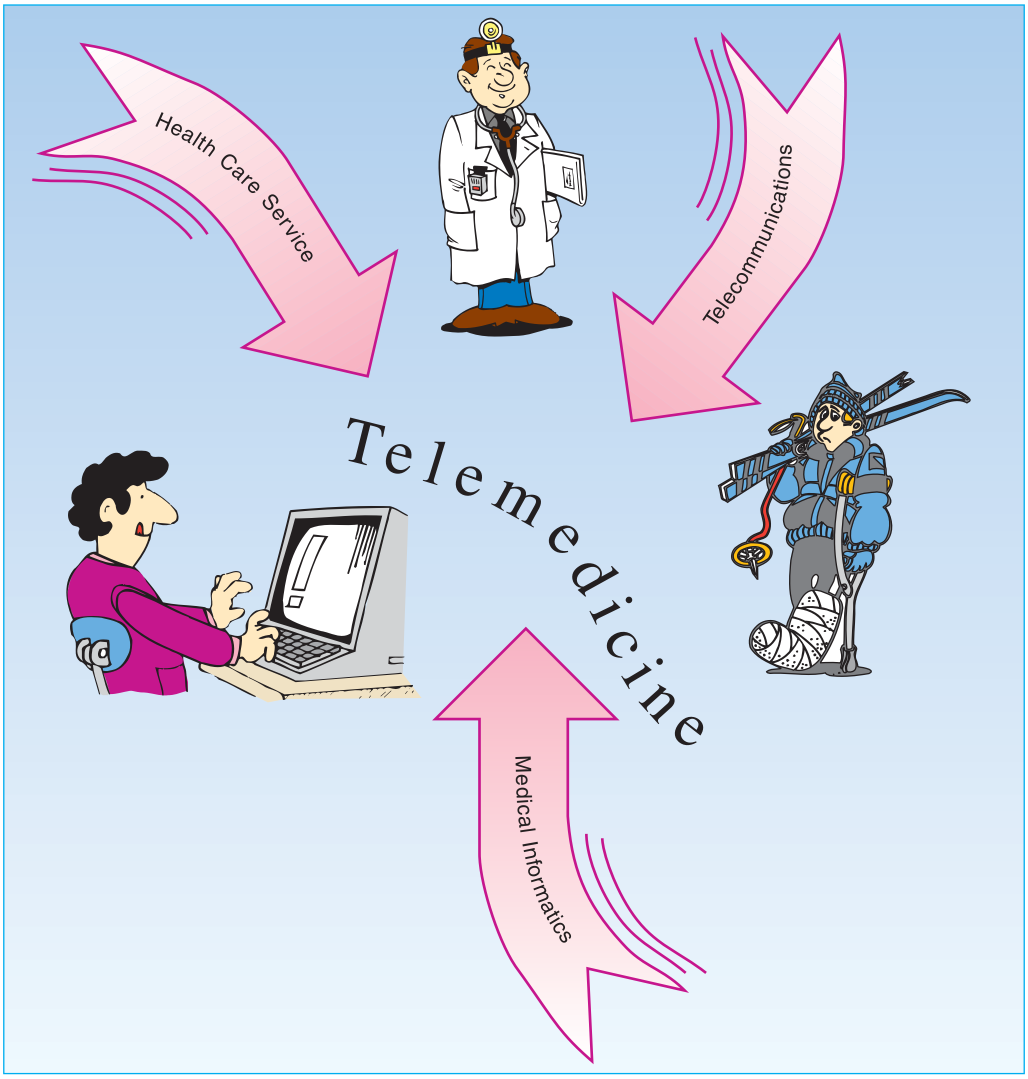


Figure 1 Telemedicine and related areas

Telecommunication

It is a misunderstanding to define telemedicine as a tele-service. Telemedicine cocerns a closer link between the telecommunication infrastructure to the health care structure, where the whole spectrum of tele-services is included. The public health service is built up as a hierarchical structure that assumes co-operation between a series of institutions locally, regionally and nationally. The need for co-operation and communication between the various levels differs. The most important communication partners for the Norwegian municipality health services are laboratories of various types (Stenvold, 1992). Specimens from patients are sent for analysis and the results are returned. Patients are dispatched from small hospitals to larger hospitals for examination by experts. In case such consultations should be replaced by remote consultations, the telecommunication network must comply with other demands than what is the case for communication with laboratories.

A central element of the telecommunication network is the transmission capacity required by the various medical applications. Communication to or between laboratories implies transmission of relatively small quantities of text. This requires only small transmission capacity. In many cases telebased consultations involve transmission of live pictures. In such cases the transmission capacity is a question of how much the quality of the pictures can be reduced and still be medically acceptable.

Another aspect is to what degree network and services allow on-line communication. In most cases the laboratory communication does not require "consultation" between GP and laboratory with regard to interpretation of the outcome of the analysis. The laboratory sends the results when they are ready and the GP studies them when it suits his work routine. Many types of remote consultations are a direct meeting between patient and physician and require two-way sound and picture communication.

As an integrated part of the working tools of the health services, telemedicine puts strong demands on telecommunication security. One aspect is that the contents of the communication must not come into the wrong hands. Another aspect is that in some cases of remote diagnosis it would be catastrophic if the connection is not obtained or if the connection breaks down.

Reliability and security of the telecommunication network is a deciding factor for which telemedical applications can be realised, for both practical and economic reasons.

Medical informatics

Medical computer science and information processing must be regarded in two connections: computer technology in medical equipment, and the more traditional internal EDP based systems within the institutions to take care of routines in the administration of health services.

Within modern health services the performance of diagnosis, treatment and rehabilitation is marked by a widespread use of medical technical equipment of a high technological level. An important element of this equipment is electronics and computer technology. However, the situation today is that this equipment to a limited extent is integrated into large internal information systems. The greater part of the equipment is a well defined entity in itself. A typical example is picture diagnosis where a computer tomograph produces pictures which are transferred to a film before they are included in the patient's case record or filed.

One problem concerning telemedicine is the necessity of a standardised interface in order to communicate information out from the equipment and through internal and external networks.

The core of electronic management of information at many health institutions are various administrative EDP systems. In the first place this concerns systems to take care of accounting, budget, payment, and material. Gradually, systems taking care of administration of patients and other activities have been put to use. The main motive for introduction of such systems is to increase the efficiency of paper work and the production of statistics. To a very small extent systems for administration, planning and medical treatment are integrated into one system.

The lack of national co-ordination has lead to a health service marked by a small degree of harmonisation concerning both hardware and functional use of medical information systems. There are many different systems which are representative of system dependence and a structure that is functionally adapted to local conditions. The lack of compatibility makes the transmission of information between the systems difficult, if not impossible. Another condition that tends to amplify these problems is the lack of standards. This concerns medical nomenclature, terminology, and classification.

The medical information systems require a high degree of security. Today the institutional systems are closed to the outside world. There is no gateway into them. This concerns protection of privacy, but even more important is to prevent "saboteurs" to enter and put the system into disorder.

The problems concerning integration, standardisation, security, and communication interface of medical information systems are relevant to telemedicine.

Health services

The organisation of health services is an important cue for telemedicine. Whether it concerns special cases or total strategies where telemedicine is involved as part of the working tools for the health services to obtain its objectives, quality control, economy, and legal conditions must be considered.

Quality control is to a high degree related to the question of whether telebased applications, e.g. for diagnosing, will produce the same results as by traditional methods. Telebased applications imply changes in the work routines, e.g. the expert does not himself guide the probe by an ultrasound examination. Another example is that the pathologist must examine the microscope pictures on an image display and not directly through the eye piece of the microscope. Another question is whether telebased transmission results in loss of details in the transmitted information, details that are important for the diagnosing. It is, of course, a basic condition for telemedicine from a medical point of view that the health services shall not be reduced in quality.

The legal aspects within medicine have become more pronounced by the focusing on the patients' rights. In the case of some medical applications it is important to clarify the responsibility towards diagnosis and treatment. A concrete example is when a patient together with a general practitioner call on a specialist via telecommunication. By traditional specialist consultation it is general practice that the specialist's instructions are valid. Thereby the specialist has the responsibility. The same will be the case by telebased specialist consultations, but this has not been tried before a court of justice. Another question is the responsibility of the supplier of the communication service in case of e.g. a breakdown of communication.

Economic considerations about telemedicine might be regarded from two levels, one is social economics and another level is related to the performance of one single health service in a particular situation. In the latter case it might be a question of replacing an existing arrangement with another one, based on telemedicine. A typical example is the arrangement with travelling medical specialists in areas where it is difficult to establish permanent specialist services. Around such arrangements is usually built up a system of economic transfer which keeps the balance for all parts involved. If this arrangement is replaced by a telebased service, the economic arrangement will not balance. Thereby clear-cut winners and losers will emerge. This might result in demand for a cost/benefit analysis and a demand for adjustment of the existing economic arrangements.

In the perspective of social economics the problem might be related to efficiency and effectiveness of the health services. The starting point is a steadily increasing gap between the expectations of the population to the health services and what the community can make accessible for health purposes. The question becomes partly how telemedicine can contribute to optimum production of health services from a given amount of resources, and partly how telemedicine can contribute to fulfil central health-political objectives. This perspective reaches far into the way of organising health services, and not least of all how co-operation and sharing of labour should be arranged.

Why telemedicine?

The superior goal of the health services is to fight sickness and promote health. Considerable resources are used for health purposes. Typical for European countries are health expenses to an amount of 7-9 % of gross national product, and it is steadily increasing. The health trade is characterised by a rapid development of medical technology. This applies to advanced instruments for diagnosing and treatment, drugs, technical remedies for nursing and biotechnology. One of the greatest challenges for the health services is to match the expectation of the population to the health services to the services which can be delivered on a large scale.

Another characteristic feature is a large production of knowledge through research and practical clinical work. The classical medical field is steadily divided into more specialised fields. At the same time the contact with other fields is extended. The health services will change character, partly because, in addition to the curative and preventive medicine, there will be more emphasis on predictive medicine based on the development in biochemistry and molecular biology.

In organising health services these development trends point towards a larger division of work duties and a centralisation of specialised functions. New professional environments and new equipment for diagnosing and treatment require comprehensive efforts of as well economic as organisational character. In the first place this will be important on a national level by the fact that some institutions will have nationwide responsibility for certain specialities imposed on them. In the long run we can se a need for greater international sharing of labour.

Another challenge for the public health services is that the population during the next ten years will be changed with regard to composition of age groups. Older people, who are the major users of health services, will considerably increase in number. In Norway the number of persons over the age of 80 has doubled since 1970. Towards the year 2020 this number will increase further. The part of the population over the age of 90 will increase considerably. In general, there will be more heavy users of the health services at the same time as there will be fewer persons in the age group which can produce health services and contribute to the economic resources needed by the health services. In addition, an increased number of handicapped and chronically sick persons is expected (St. meld. (White Paper) No. 41, 1987-88).

As an answer to the increasing cost of running the institutions of the health services (hospitals, nursing homes), we will have a change towards greater emphasis on primary health service. In Norway we are already phasing over to more day care activity where staying time in hospital is reduced and in the nursing sector a transition to homebased care.

In light of these challenges telemedicine must be assessed as a tool for more efficient exploitation of available resources. Telecommunication will never replace the physician or other health workers involved in a patient relation. Instead, it gives a possibility of increasing the integration between various health services and in this way contributes to better care directed towards the patient.

From technological curiosity to economical benefit

The idea of using telecommunication for medical purposes is as old as the spread of the telecommunication means. Soon after the invention of the telephone experiments were made to transfer heart and lung sounds to a skilled specialist who could give an opinion of the state of the organ. The inventor of the electrocardiograph, Wilhelm Einthoven, started experiments with remote consultations via the telephone network (Einthoven, 1906). Also in Norway such possibilities were utilised. Haukeland Hospital established in the 1920's a service where ships at sea could consult physicians in hospitals via Bergen Radio in case of accident and sickness. It has ben said that the physicians not only contributed with diagnoses and proposals for treatment, but also complicated surgical operations were performed by help of instructions via radio (Rafto, 1955).

During the 1950's and 60's many individual experiments with medical services were carried out on the basis of telecommunication. Often it was enthusiasts with medical background who saw the possibilities as the teletechnology gradually developed. We may safely assert that those experiments were mainly directed towards the technology, even if medical and organising matters were on the agenda. The equipment used was poorly adapted to the services to be practised. The cost might be so high that the data obtained could not be generalised and lead to safe conclusions (Bashur and Lovett, 1977).

Gradually the starting point for development of telemedicine changed towards the solution of concrete medical problems. Such a field was supervision of physiological functions of crews in space ships (Pool, Stonesifter and Balasco, 1975). Another field was improvement of primary health services in areas with scattered population (Fuchs, 1979, Dunn and Higgins, 1984). Telecommunication was put to use for remote consultations and remote diagnoses and for distance education of medical personnel at remote locations.

With the linking up of teletechnology with data technology the horizon of telemedicine was appreciably extended in the 1980's. Within medical computer science a series of data programs and systems were developed of both administrative and medical varieties. Even if they were intended to take care of internal tasks in institutions, they laid the foundation for new and also improved older telemedical methods. As an example, digital picture processing has now obtained a central place in several telemedical applications.

However, at the same time it may be asserted that while the health services have been progressive in adaptation of advanced medical technology, far less attention has been focused on the use of telecommunication and information processing. Introduction and acceptance of this type of technology have been slower within the health services than in several other fields (AIM, 1992). This concerns the more administrative sides for personnel, institutions, and patients, but especially within medical treatment. This is in contrast to the fact that the health services have been very information intensive at all levels. As an example a hospital bed in Europe represents a yearly production of X-rays amounting to an average of 1 Gigabyte (France and Santucci, 1991).

During the 1990's the strongest driving force for development of telemedicine is the economic dimension. Viewed from one angle this is due to the challenges facing the health services, where greater efficiency in performing the health services can moderate the conflict between access to resources and the expectations and demands of the population. Information technology is regarded as an important tool for increased efficiency (World Health Organisation, 1988, Arthur D Little, 1992). On the other hand the health services in this perspective represent a considerable market for information technology (France and Santucci, 1991).

In Europe the economic driving force is clearly demonstrated by the two development programmes under EC direction concerning health services and telecommunication, by AIM (Advanced Informatics in Medicine) and the RACE (Research and development in Advanced Communication in Europe) project TELEMED. In both programmes the participants are a composition of research institutions, medical institutions and industry concerned with information technology.

In the TELEMED project the perspective is to find the problems emerging when medical experts communicate through a broadband network. Based on trials and experiments the experience is transformed into technical specifications regarding equipment for telecommunications and terminals. The objective is to produce commercial systems adapted to co-operation between medical experts within diagnosis and therapeutics (TELEMED 1991).

In the AIM programme objectives are clearly expressed to promote a more efficient co-operation within the health services through development of tools, techniques and practice about medical informatics and telecommunication with a common European foundation. A further objective is to prepare the European market and strengthen the competitive force of European industry within this field (AIM 1992).

Telemedicine in Norway

The Norwegian public health services are decentralised and based on the principle of treatment on the lowest efficient level of care. The primary health service is the responsibility of the municipalities while the counties are responsible for the hospitals. Each county has a central hospital and a varying number of local hospitals. The country is divided into five health regions. Each region has a regional hospital with the responsibility of taking care of special medical competence and treatment. On a national level there are also some hospitals to take care of particularly rare or complicated types of illnesses.

The hierarchical structure assumes co-operation between a series of units. This implies transfer of large amounts of information between the various levels. In the same way as for the rest of Europe, electronic processing of information is largely tied to administrative routines inside the institutions. The public health service lacks an integrated electronic information system. It is rightly asserted in several connections that management of information is lagging behind in time in relation to the demands created by the development of treatment of patients (Health region 3, 1988).

However, in one field the work with better resource utilisation through integrated information exchange has progressed; i.e. medical emergency reporting services. By law a common national emergency reporting service is to be established, which co-ordinates communication readiness for various medical levels and co-operating services such as police and fire service (Odelstings Prop. No. 26, 1988-89). The core of the communication system is a communication centre for acute medical needs on a county level, where the various telephone and radio services can be exploited for co-operation.

An example of a more classical incitement to telemedical thinking is the project "Telematics in the health service of Finnmark" When the idea was launched in 1986, the starting point was the lacking coverage of health personnel, long distances, and in general poorer health services in this county than in the rest of the country. The main ingredients of the project (Andersen, 1992) were video conferences for remote diagnoses, education and professional meeting activities between the county hospital and the University Hospital of Tromsø. The experience gained was satisfactory to a degree that the local health authorities have decided to include telematics in their strategic plans and to strengthen the co-operation between health services within and outside the county.

Telemedicine in North Norway

The largest effort within telemedicine in Norway is a project under the direction of Norwegian Telecom. Following the good result from an experiment with transmission of ultrasound pictures through the telecommunication network between the local health services at Jevnaker and Ullevål Hospital in Oslo (Andersen and Nordby, 1988), the project "Telemedicine in North Norway" was established. There are two reasons why the project was located in the northern part of Norway: One is the geography of that part of the country, scattered population and in many places poor heath services, especially for medical specialist services. Secondly, this area has a well developed infrastructure of telecommunication, especially networks for video conferences and video transmission (Nymo, 1989).

From the outset the project tried to incorporate the three dimensions of telemedicine, namely medical informatics, health services, and telecommunication. This has made the approach to the project rather complex. It ranges from development of technical equipment through participation in international standardisation work to assessment of questionnaires concerning the satisfaction of the patients with the telemedical services. The superior objective was to arrive at medically secure, technical, and organising solutions for the Norwegian health services. A joint approach to the problems was field trials where an approach crossing professional borders was taken care of through co-operation between the Norwegian Telecom Research (NTR), other research institutions, individuals and institutions in the health services. The most important collaborator for the field trials has been the University Hospital of Tromsø (RiTø).

The field trials can be characterised as explosive. Ideas around techniques and procedures have been developed in a real situation, and these have been further developed into equipment and routines. Thus systems for telepathology and teleradiology have been developed, systems which have been commercialised and which are currently in operation. Tests have been made within this area which confirm the fact that the systems render a satisfactory quality of medical services.

The field trials have also had a considerable educational effect. The trials have

Figure 2 Studios for video conferences

concretised the possibilities offered by information technology for the health services. Various participants from the health services have from their own point of view put information technology on the agenda as a solution of challenges in the health services. Thus, a structured effort, initiated by the Department of Health and Social Affairs, is in progress to standardise medical information transfer within the health services. A special centre for competence concerning information technology in the health services is established. This centre has, among other things, led negotiations for a framework agreement for services for the health sector based on electronic mailboxes. One triggering factor has been the work with transmission of assessment of samples from the laboratory to the recipient of the information. In the county of Nordland telemedicine is included by the authorities as an important tool to achieve a superior objective of more efficient hospital services. At last there is a clear tendency that the various medical disciplines take the initiative for new telemedical experiments. A special nationwide function within telemedicine is established at RiTø in order to take care of national competence in the field and to provide efficient use of telemedicine.

Simplified, the work span of the project can be divided into two axes:

- remote diagnoses and remote consultations
- electronic transfer and collection of information.

Remote diagnosing can be described as a medical practice where a medical specialist diagnoses a patient based on teletransmitted information about the patient. Particular focus has been placed on information through pictures. Direct contact between the medical specialist and the patient depends on whether the specialist is in a clinical field (e.g. skin, psychiatry) or in a laboratory field (e.g. pathology, radiology). Remote consultation implies that a medical specialist confers with another medical specialist about particu-

Medical field	Telemedical services	Aspects of telecommunication
Pathology	- Remote diagnosis frozen picture	- Remote control of microscope/ Interactive video/audio communication
	- Clinical pathological conference	- Interactive video/audio communication Still pictures from microscope
	- Remote diagnosis electro-microscopic pictures of muscular tissue	- Digitised still pictures/Data communication
Radiology	- Remote diagnosis ordinary X-ray pictures	- Digitised still pictures/Data communication
	- Remote diagnosis computer-tomographic pictures	- Data communication
	- Remote diagnosis ultrasonic pictures	- Interactive video/audio communication
	- Radiological remote consultations	- Interactive video/audio communication
Dermatology	- Remote diagnosis of patient	- Interactive video/audio communication Still pictures
Otorhinolaryngology	- Remote diagnosis of patient	- Interactive video/audio communication Video pictures from endoscope
Microbiology	- Remote diagnosis of bacterial growth	- Digitised still pictures/Data communication
Psychiatry	- Psychotherapy - Consultation with local health workers	- Interactive video/audio communication - Interactive video/audio communication
Gastroenterology	- Education	- Digitised still pictures/ Interactive video/audio communication
	- Expert consultations	
Cardiology	- Education - Remote diagnosis of patient	- Interactive video/audio communication - Video pictures from echo dopler

Figure 3 Table of remote consultations

lar difficult cases, interpretation of findings, etc.

The trials with remote consultation and remote diagnoses included in the project can be structured according to medical fields, the type of telemedical services involved and the various aspects of telecommunication.

In addition to the medical services practised during remote consultations, matters concerning education, instructions and co-operation between health institutions are also kept in mind.

Within electronic exchange and collection of information the following subjects are dealt with:

1 Mapping of communication structures

The objective of mapping information handling and communication within various disciplines and various levels in the health services is to introduce information models as a foundation for the demands to be complied with for the aids to information and communication. Mainly, it concerns two types of information: patient information (patient case record, applications, requisitions, etc.) and operational information (economy, circulars, statistics, etc.).

2 Standardising medical document exchange

Information systems in the health services must be characterised as isolated systems in view of their limited possibility of interplay. Transport of information between participants and further treatment by the recipient is thereby limited. The standardisation work is intended to harmonise the information handled by various data processing systems so that information from several systems can be integrated and compared. The second demand is a form of electronic handling that is accessible to all the communication participants. The project has participated in national as well as international standardisation work.

3 Sample analysis and laboratory communication

To obtain practical experience with electronic information exchange, a system is made for the handling of sample analysis results from a clinical-chemical laboratory to the recipient of the information. Clinical-chemical laboratories are some of the most important communication collaborators for the primary health services. It is a stated requirement for more efficient laboratory communication.

Experience from the work with laboratory communication is closely related to verification of specifications evolved through the standardisation work.

4 Electronic access to medical knowledge and experience

Large amounts of knowledge are generated through research and practical work within medicine and related professional fields. The quality of health services depends on the professional qualifications of the health personnel. It is not a trifling challenge to keep oneself professionally updated. There are several electronic sources for medical information in the form of facts databases, literature databases and electronic conferences for particular professional fields. The project has been working with a system for medical information services (MEDIS) where emphasis is put upon making access to the information sources as simple as possible for the users. Among other things, this is achieved by making a uniform interface to the various information sources.

The challenge

As data technology has taken over from electromechanical equipment in telecommunication, the horizon for exploitation has been considerably extended. However, a condition is that data technology has a widespread use in the community, where it is used to systematise, store, file, and treat information. When this is the case, we have a technical foundation for collection, reception and treatment of information by help of telecommunication. Thereby, new fields for exploitation are found. It may be a long road from the fact that technical foundations exist to the realisation of new fields for exploitation. This is related to the need for social changes in the form of organising, work routines and roles. It is a challenge for the health services, the supplier of telecommunication and for the producers of medical technical equipment and medical information systems.

The challenges facing the health services make it necessary to try new methods for practising the services. The main problem in Europe, and even more so in the USA, is the rising cost involved with health services. Better utilisation of resources through extensive co-operation between participants is a more realistic solution than increased share of the society's expenses for health purposes. In this light more focus should be directed towards telemedicine in a fully organising and strategic perspective.

This implies that the field of telemedicine changes from a local or regional starting point with focus on loose concrete problems, giving poor coverage of specialist services in rural areas, to the making of a telemedical infrastructure that matches the general infrastructure of the health services. The structure can be based on the existing work sharing between the levels of the health services, but it must have enough flexibility to realise new patterns of co-operation. For example, there are many reasons why greater emphasis should be put on homebased rather than institutional care for the growing number of persons in need of nursing. Thus, the infrastructure must also include homebased care. As the sub-specialisation continues, it might be impossible for small nations to establish lasting professional environments on a national basis. Professional environments in the form of networks across national borders is one way to organise such services.

There are a great number of applications of telecommunication within the health services. Some are more important than others. An American investigation (Little, 1992) calculated the potential for the yearly reduction of costs for some application groups. The application group found to give the highest reduction of costs is electronic handling and transport of patient information. The investigation calculated that in the USA it is possible to save USD 30 billions (30,000,000,000) yearly by employing such applications. The potential for cost reduction by systematic use of video conferences for remote consultation and education was estimated to a yearly amount of USD 200 millions.

Systematic exploitation of applications assumes flexible possibilities for telecommunications. In today's pilot projects and other work within telemedicine, by and large the applications are realised in general teleservices in different networks. For some purposes this does not result in the best cost optimum solutions. At the same time users' choice of type of communication is limited. For many medical purposes where pictures are part of the information basis, there is a need to choose transmission rate according to a particular situation. In an acute case there are other demands for speed of connection and transmission than for routine examinations.

ISDN seems to give the necessary flexibility by using a combination of basic access and primary rate access. The primary health services' need for communication both to local hospitals and in the homebased care, basic access offering two 64 kbit/s channels (B-channels) in the great majority of cases give sufficient capacity for the applications we see today. For telemedical services between hospitals the flexibility will be secured by primary rate access (30 B-channels). In a complete "health network" it is also necessary to integrate mobile communication, especially in cases of catastrophes.

References

Advanced Informatics in Medicine (AIM), Commission of the European Communities. *Supplement Application of Telecommunications of Health Care Telemedicine.* AI 1685, 1990.

AIM'92. *Research and technology development on telematics systems in health care.* Brussels, Commission of the European Communities, 1992.

Andersen, K A. *Videokonferanser i helsevesenet, erfaringer fra to års bruk i Troms og Finnmark.* Kjeller, Norwegian Telecom Research. (NTR paper N3/92.)

Anderson, D, Nordby, K. *Video telephony revisited: The development and test of a broadband multifunction terminal.* Lecture held at the 12th International Symposium on Human Factors in Telecommunication, the Hague, 1988.

Bashur, R L, Lovett, J. Assessment of telemedicine: Results of the initial experience. *Aviation, Space and Environmental Medicine,* 177, 65-70.

Bird, K T. Telemedicine; concept and practice. In: R L Bashur, P A Armstrong and Z I Youssef (eds.), *Telemedicine; explorations in the use of telecommunication in health care.* Springfield, Illinois, Thomas, 1975.

Conrath, D W et al. *Evaluating telecommunications technology in medicine.* Dedham, Massachusetts, Artech House, 1983.

Dunn, E V, Higgins, C A. Telemedicine in Canada; an overview. *Dimensions,* 68, 16-18, 1984.

Einthoven, W. Het telecardiogram. *Nederlens Tijdschrift voor Geeneskunde,* 50, 1517-1547, 1906..

Fuchs, M. Provider attitudes toward STARPACH: a telemedicine project on the Papago Reservation. *Medical Care,* 17, 59-68, 1979.

Helseregion 3: telemedisinutvalget, Bergen, 1988.

Little, A D. Telecommunications: *Can it help solve America's health care problems?* Cambridge, Massachusetts, Arthur D Little, 1992.

Nymo, B J. Telemedisin i Nord-Norge. *Telektronikk,* 85, 260-263, 1989.

Ot. prp 26(1988-89): *Om telekommunikasjoner i helsetjenesten, herunder medisinsk nødmeldetjeneste.*

Pool, S L, Stonesifter, J C, Belasco, N. *Application of telemedicine systems in future manned space flight.* Lecture held at the 2nd Telemedicine Workshop, Tuscon, Arizona, 1975.

Rafto, T. *Telegrafverkets historie 1855-1955.* Bergen, Grieg, 1955.

Roger France, F H, Santucci, G (eds.). *Perspectives of information processing in medical applications. Strategic issues, requirements and options for the European Community.* Brussels, Springer-Verlag, 1991.

Stenvold, L A. *Kommunikasjonsstrukturer i helsevesenet, en gjennomgang av kartleggingsforsøk i telemedisinprosjektet.* Kjeller, Norwegian Telecom Research. (NTR report R7/92.)

St.mld. 41(1987-88). *Helsepolitikken mot år 2000. Nasjonal helseplan.*

TELEMED-Project: *Annual project report 1991.* Berlin, DETECON GmbH, 1991.

World Health Organisation. *Informatics and telematics in health. Present and potential uses.* Geneva, WHO, 1988.

Telemedicine services integrated into a health care network
– Analysis of communication needs in a regional health care system

BY SIGURD FROM, LILLY ANN STENVOLD AND THORE DANIELSEN

621.39:61

1 Introduction

Since 1988 Norwegian Telecom Research has initiated and developed several telemedicine applications. Telemedicine services supported by the applications range from exchange of clinical chemistry laboratory results to interactive radiology consultations. All the applications have a common objective to improve efficiency and quality of the health care.

This paper defines a set of objectives for introducing telemedicine in the Norwegian health sector. It also outlines the structure of the regional health care system in Norway which will be the scope for this paper. Communication needs within a typical health care region is analysed. Later, telemedicine applications which meet these needs are described by scenarios. Finally, the applications are integrated to form an example of a regional health care network.

2 The objectives of telemedicine

One of the basic ideas of telemedicine can be expressed by the saying: "Move the information, not the patient". When a patient needs to consult a specialist, information about the patient could be obtained locally and exchanged through a network to a specialist. In many situations this can replace transporting the patient or the specialist to a given location. This exchange of information for medical diagnosing and treatment is a basic concept of telemedicine.

In general, we may use the following functional definition of telemedicine (AIM, 1990):

> *The investigation, monitoring and management of patients and the education of patients and staff using systems which allow ready access to expert advice and patient information no matter where the patient or relevant information is located.*

Telemedicine may therefore be seen as all situations where information is exchanged electronically between health care parties that collaborate in treating patients.

In 1992 the Norwegian Ministry of Health defined 5 information technology (IT) objectives for the Norwegian health care sector (Hidas, 1992):

IT1 Improve service and quality of the health services

IT2 Improve productivity and efficiency in the health sector

IT3 When meeting IT1 and IT2 the aspects of security of privacy for the individual patient should be considered

IT4 Use the opportunities of IT to distribute information to the general public and the health care professionals and to increase the level of knowledge

IT5 Improve working conditions and personal planning for health care professionals.

These objectives relate to the whole of the health sector, not only telemedicine. However, those objectives directly related to patient care may be used to form similar objectives of telemedicine. By combing IT1-5 and the previous definition of telemedicine we have defined the following 7 objectives of telemedicine:

TM1 Patients should be treated as close to their homes as possible (IT1).

TM2 Medical expertise should be equally available independent of where the patient lives (IT1).

TM3 The quality of medical decisions should be improved by making existing information about patients more easily available (IT1).

TM4 Patients should get more information and better service (IT1).

TM5 The health services should improve efficiency and productivity by reducing unnecessary administrative work such as retyping information already existing in electronic form and by distributing tasks between health care institutions and health care personnel (IT2).

TM6 All exchange of information needs to take into account the aspects of security of privacy for the individual patient (IT3).

TM7 Medical knowledge should be more easily accessible (IT 4).

3 Communication scenarios

The following section explains the method used to describe telemedicine applications in brief. Telemedicine applications support the process of communicating between groups of health care workers. Such group processes have earlier been described in Computer Support Co-operative Work (CSCW). We have chosen to use the AMIGO Activity Model (Pankoke-Babatz, 1988) combined with some concepts from the Open-EDI work (SIS-ITS) as a basis for our method.

In our context a *telemedical application* is the total solution implemented to support the need for a telemedicine service. It includes both software and hardware at all party sites, and the communication between them. A *telemedical service* is a medical service provided by one party to one or more other parties through applications. A telemedicine application may support one or more telemedicine services.

The overall picture of the activity supported by a telemedicine application is described in a *communication scenario*. As for theatres, a scenario consists of a set of *roles* played by a number of persons or role players. We will focus on the roles participating in the communication process. An actual role player is called the *communicator*. In health care the communicator is typically a health care party, i.e. a health care professional such as a physician, a nurse or a health care institution such as a laboratory. One role is associated with one or more *functions* that describe the actual work done, the behaviour of the role. One role may send *messages* to one or more other roles which trigger a function in that role to be performed. In the situation where requester and provider roles are involved in the scenario, a *service* is defined as the function performed by the provider role upon request from the request-role. Figure 1 illustrates the relationship between the various concepts.

The following gives a summary of the concepts defined:

- Communication scenario: A collection of matching roles grouped together to support an activity with a given purpose.
- Role: The definition of a communicative role of communicators in a scenario describing the specific behaviour of the communicator. Examples of roles are: laboratory requester and laboratory service provider. There may exist a number of communicators attached to any role, and any commu-

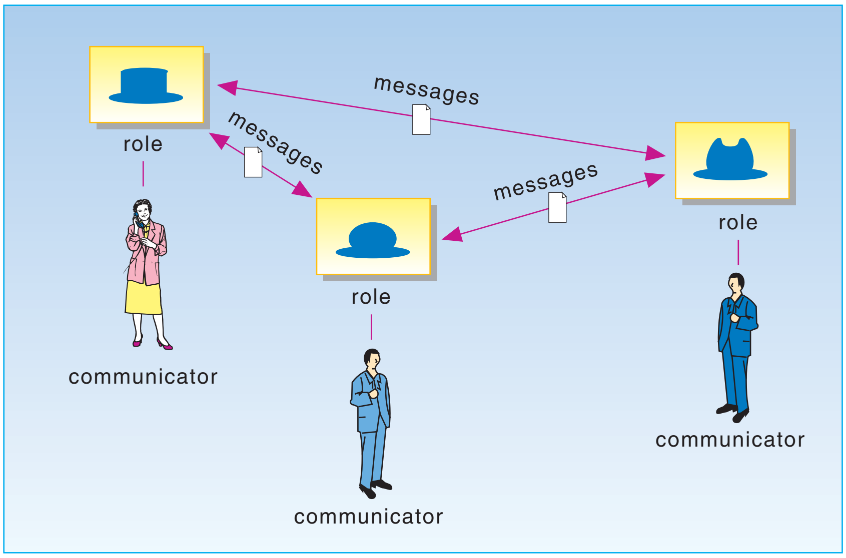


Figure 1 The relationships between the various concepts

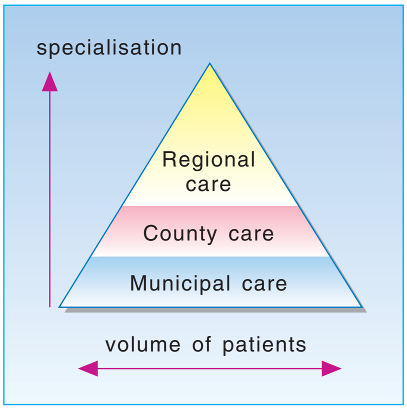


Figure 2 Specialisation and volume of patients in regional health care system

nicator may act in a number of roles conversely.

- Communicator: The parties of any group process are described as communicators, i.e. persons or processes capable of taking part in communication. In the health service communicators are typical health care parties, i.e. health care personnel or health care organisations.
- Message: A logical information unit exchanged between two roles in a scenario. All messages have a given purpose which may trigger a dedicated function in the receiving role.
- Function: A unit of behaviour performed by a role. A role may consist of one or more functions.
- Service: The function performed by a role upon request from another role.

4 The regional health care system of Norway

The need for telemedical services is based on the need to communicate between parties in the health care. This chapter will give an overview of the structure of the regional health care system in Norway and identify some of the parties involved. The parties, as will be seen later on in this text, are potential communicators in telemedicine applications.

4.1 The organisational structure

The health care sector comprises different general and specialised health services. The tasks include preventive practice, medical treatment, rehabilitation, and general health care.

In Norway, the health care sector is organised on the basis of geography and the density of the population. The health care services are divided into three levels:

- municipal care
- county care
- regional care.

The municipal care is the most general level, consisting of services for small population units, such as primary care services. The county care comprises more specialised medical services, such as institutions and small hospitals. The regional care is the most specialised level, and includes regional hospitals. Norway is divided into five health regions, each with one regional hospital. Often the term primary health care is used for the municipal care and secondary care for county and regional care.

Each party in the regional health care system represents different levels of specialisation, from the most general medical services in the municipal care to the most specialised medical services in regional care, see figure 2. An important principle for this organisation of the health care sector is called “lowest efficient care level - LECL”. This implies the patients to be taken care of at the lowest level that can provide sufficient care, and at a level specified by the patients’ medical needs. This has impact on resource utilisation and on the planning of capacity for each level of the health care services.

The information exchanged in the health care services can be divided into functional categories based on the context of information:

- patient information
- medical knowledge and experience
- administrative information.

Patient information comprises information that can be connected to a specific patient, e.g. in the medical record. Aggregated patient information may be used to form statistic information. Medical knowledge and experience form the general basis for taking care of the patient. Administrative information includes information about resources such as personnel, finances, and equipment. Administrative information is usually not a part of telemedicine applications.

4.2 Parties in the health care system

4.2.1 Patient

A patient is described as a person who receives medical examination, treatment, guidance, or care from the health care services. The contact between the patient and the health care services initiates the process of care. The patient is therefore an important party in the health care system.

4.2.2 Health care parties

The health care parties are the actors, those who provide care, in the health care system. Health care parties are either persons, i.e. health care personnel such as a physicians, nurses, or technicians, or they may be organisations or part of organisations such as a surgical department of a hospital.

The regional health care system comprises a number of different health care parties who in their roles perform different functions within the system. For the purpose of later identifying the needs for telemedicine services in this system we describe the following health care parties:

- General practitioner (GP)
- Pharmacy
- Medical library
- Hospital
- Surgical department of a hospital
- Dermatological department of a hospital
- Pathological department of a hospital
- Radiological department of a hospital
- Clinical chemistry laboratory.

The general practitioner (GP) works in municipal health care. The patient gets in touch with the GP for medical care at a general level or to be referred to other parts of the health services. A consultation describes one function of a general practitioner, and may include medical investigations, medical treatment, or supervising of the patient.

The pharmacy makes up prescriptions from a medical practitioner for a specific patient.

Medical libraries contain references to literature and collections of books, articles and databases with general medical information and medical experience. This is an important source for medical knowledge, especially for medical practitioners.

Hospitals carry out the more specialised tasks within the health service. The main units in hospitals are departments, which may be grouped into three categories according to their functions:

- clinical departments
- medical service departments
- non-medical service departments.

The medical functions are taken care of by clinical departments and medical service departments. Clinical departments comprise outpatients' departments and wards. One example of a clinical ward is the dermatological department.

Medical service department is a term for the medical departments which do not have their own wards, but whose main task it is to assist the clinical departments in special examinations and treatment of the patients. Some examples of medical service departments are anaesthesia departments, X-ray departments, and various medical laboratories. There are a number of laboratories aiming at medical specialisation. The services offered are related to medical speciality, equipment, as well as personnel. The laboratory may for example receive requests and samples, obtain samples directly from the patient, perform investigations and examinations, interpret results and convey replies in the form of reports. Examples of laboratories are clinical chemistry laboratory, microbiological department, and pathological department. Parties outside the hospital, e.g. a GP, may also use these services.

The non-medical service departments at a hospital comprise administrative and financial services, technical department, kitchen, laundry, and such like. These departments are not normally included as parties in telemedical applications.

4.3 Message types exchanged in the health care services

Information exchanged between communicators who act in the health service is in the form of documents or messages. The reason for the exchange of a message is the aim of having a service performed, obtaining information, etc. Some examples of messages are:

- letters to patients
- letters of referral
- discharge letters
- requests
- prescriptions
- reports.

Letters from health care parties to patients may contain information about investigation results, appointment for consultations, and so on.

A letter of referral means that the patient is referred to another health care authority. Referral may be made to admittance in an institution, to an outpatients' consultation, or to various medical experts. A letter of referral is sent from a physician to a health care party.

A discharge letter is a concluding summary concerning a referred patient. It is forwarded to the referring physician. Usually the discharge letter is exchanged between physicians, and it should give the referring doctor sufficient information on what is done with the patient and on plans for further treatment.

A request is used for ordering or directing something. Examinations may be ordered, like laboratory examinations, medical treatment (e.g. physiotherapy), goods, and services (e.g. drugs, aids, and transportation services). A prescription of drugs is a permit or order for a pharmacy to sell a specified drug to a specific patient.

A report is the feedback to applicant's request. The applicant may e.g. receive the results in the form of laboratory reports. For some requests there may be no need for a feedback, e.g. a prescription of drugs.

Messages are made up from information elements, e.g. a referral of a patient to a hospital consists of information on the patient, information on referring party, information on receiving party, and information on reason for referral. Messages may also be categorised according to which medium is used. A message may e.g. consist of text data, picture data, graphic data, and sound.

5 Needs for telemedicine services

Each party in the regional health care system communicates with the other parties in a region. The network of communication between the parties may be analysed by identifying each party's need for telemedicine services. In this paper we have selected three parties for a further analysis: a general practitioner in the primary care, a surgical department at a hospital, and a radiological department at a hospital, the latter two in the secondary care. For this paper an inexhaustible list of the needs for telemedicine service for those parties has been set up. The needs are presented in the order of how frequent the services are normally used.

The general practitioner needs

- to prescribe drugs
- to inform patients
- to request laboratory investigations
- to request diagnostic investigations

- to refer patients to hospital
- to look up medical information
- to interactively discuss medical issues with an expert.

The surgical department of a hospital needs

- to request laboratory investigations
- to request other diagnostic investigations
- to have conferences with experts in pathology.

The radiological department of a hospital needs

- to get radiological images interpreted by specialists.

6 Telemedicine applications

The needs presented above must be met by solutions for how telemedicine services can be provided. This chapter will use the method introduced in chapter 3 and outline a set of communication scenarios for telemedicine applications. Each scenario will support one or more of the needs identified in the previous chapter. The scenarios will be presented according to the sequence of the needs in the previous chapter.

6.1 Prescription of drugs

A number of medical services may be requested from outside a GP office. The need to prescribe drugs from a pharmacy is an example of such a service. When prescribing drugs, two parties are involved: the role requesting physician and a provider of drugs. The requesting physician issues a message, a prescription, to the provider of the drugs. The patient collects the drugs at the provider. In our example a GP fills the role of a requesting physician and the pharmacy fills the role of the provider.

Electronic prescription of drugs should increase the service for the patient allowing the drugs to be prepared at the pharmacy before the patient arrives, which means better service to the patient (TM4). In addition, electronic information exchange allows for higher security than today's paper based system (TM6).

6.2 Information to patient distribution

An important part of the care process is for the physician to keep the patient informed about the treatment and the practical arrangement around it. This results in a lot of letters to the patient and adds a heavy administrative overhead.

Electronic submission of these letters may reduce this overhead. Even though the patient herself cannot receive documents electronically, messages may be sent from the care provider by electronic mail to a conversion unit which prints out the messages, puts them into envelopes and mails them to the patient by ordinary post.

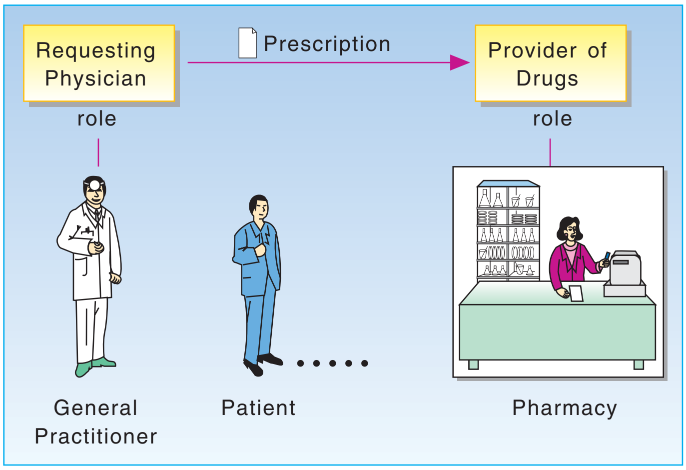


Figure 3 Prescription of drugs

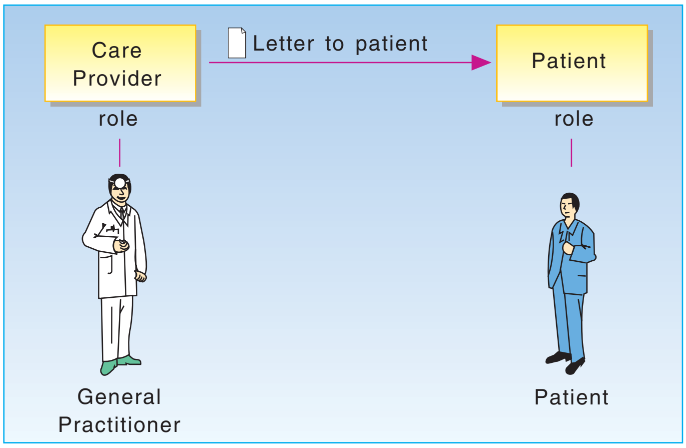


Figure 4 Information to patient distribution

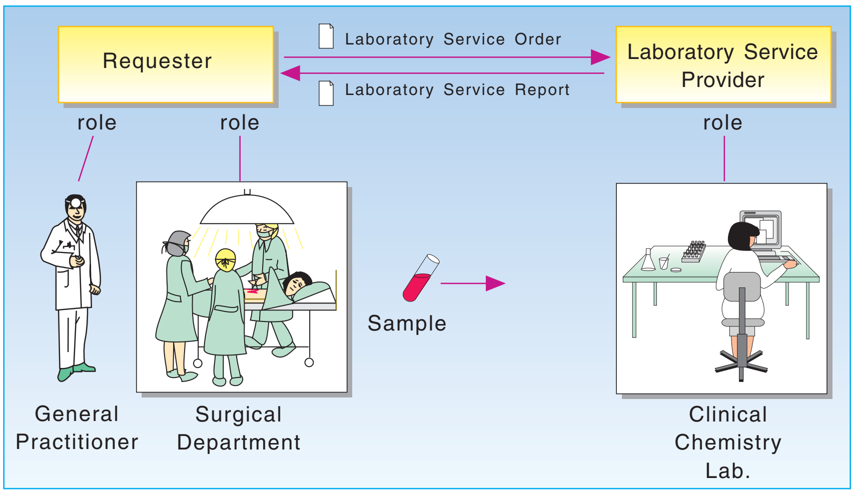


Figure 5 Request of clinical chemistry tests

The roles involved in this scenario are the care provider, which issues the letter, and the patient. In our example the role of a care provider is filled by a GP and the patient role is filled by the patient party.

The “information to patient” service should improve the quality of the information that the patient receives (TM4). However, the highest benefits are probably related to the reduction in administrative overhead (TM5).

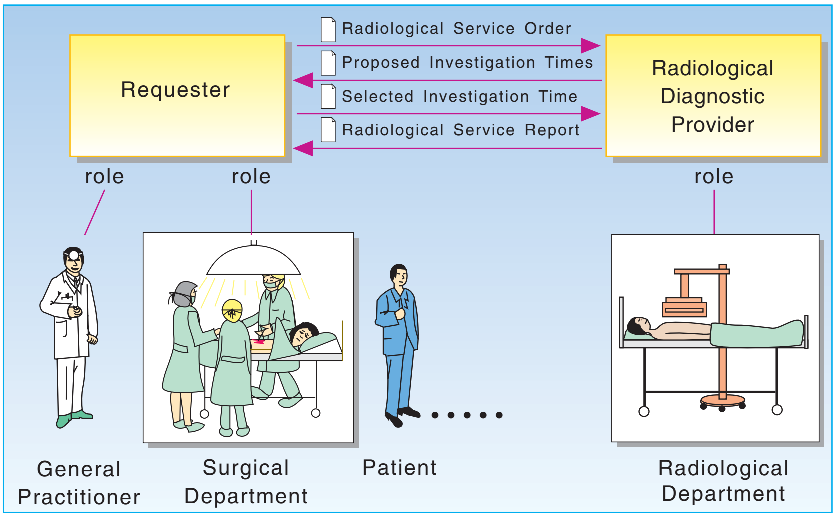


Figure 6 Referral to radiological investigation

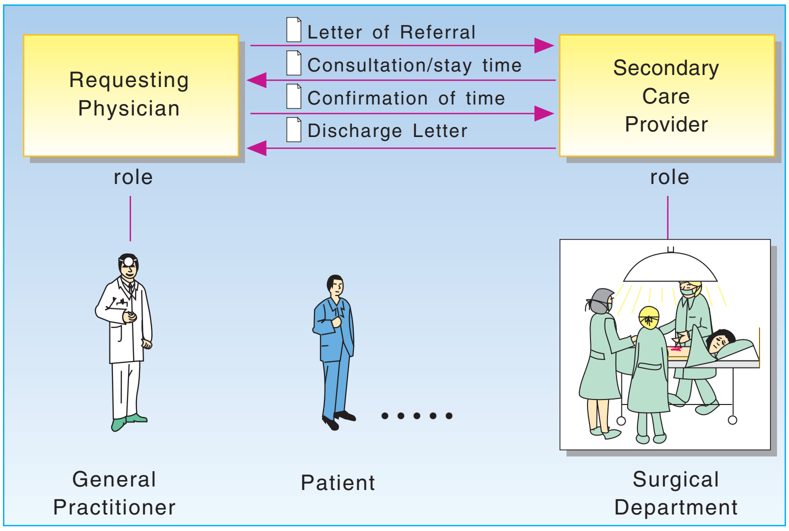


Figure 7 Referral of patient to hospital

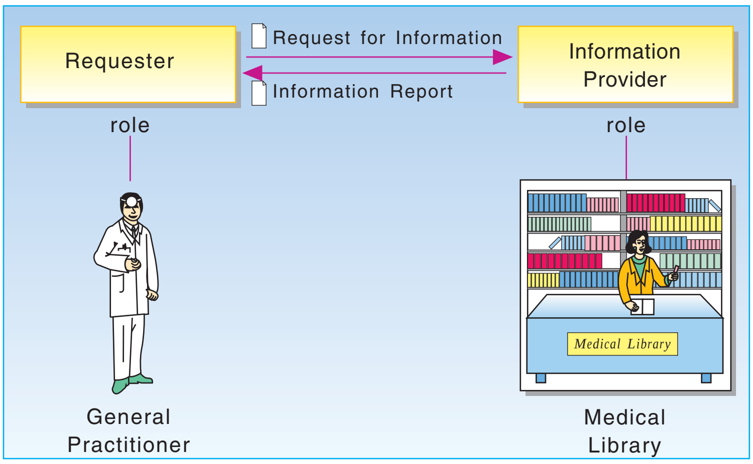


Figure 8 Medical information service (Medis)

6.3 Requesting of clinical chemistry investigation

Diagnostic services play an important role in arriving at the patient’s diagnosis. One category of the diagnostic services is the laboratory services where a sample, e.g. a blood sample, is taken from the patient and sent to a laboratory. By use of electronic communication both the order and the report may be exchanged electronically between the requester and the laboratory service provider.

Our example illustrates the situation where a GP in the role of a requester requests services from a clinical chemistry laboratory in the role of laboratory service provider. A laboratory service order is exchanged from the requester to the provider and a laboratory service report is returned.

The same scenario may also be used in the secondary care between a surgical department of a hospital and the hospital’s internal laboratory.

Electronic laboratory communication improves the quality of the information exchange which should result in higher quality on the medical decision (TM3). The effort of retyping the information for the receiving party is removed, which should result in a more efficient service (TM5). The possibility of encrypting electronic data allows for a higher level of security (TM6).

6.4 Referral to radiological investigation

Radiological investigations often require special imaging equipment and medical expertise not normally available in a small GP office. The patient must then be referred to a radiological unit. This scenario includes the roles requester and the radiological diagnostic provider. The requester sends a radiological service order to the provider. The provider evaluates the request, and proposes alternative investigation times. The requester and the patient select one time of appointment. The patient sees the provider and the investigation is performed. After the investigation the provider returns a radiological service report to the requester.

In our example the role requester is either a GP or the department of surgery of a hospital. The provider is a radiological department of a hospital.

Electronic exchange of patient information should allow for more easily available information to the decision maker.

This should result in better care (TM3). The possibility for the patient to select the time for investigation gives more flexibility for the patient and results in better service (TM4). Again a higher security level on the transmission compared to ordinary mail is foreseen (TM6).

6.5 Referral of patient to hospital

The GP may need to refer the patient to a specialist. The specialist is part of the secondary care. He may be an independent consultant or part of a large organisation such as a hospital. In this scenario the roles requesting physician and secondary care provider are made general to cover all kinds of referrals to secondary care.

The requesting physician sends a referral letter to the provider. The patient is booked in by the provider and he returns the consultation/stay time. The appointment time is either confirmed or turned down by the patient. The patient sees the provider, which may imply a long travel. After the consultation/stay a discharge letter containing the a summary of the important diagnosis, treatments, and further follow-up is sent from the provider to the requester.

In our example the requesting physician is a GP and the provider is a surgical department at a hospital.

The same telemedicine objectives (TM3, TM4 and TM6) are met by this category as for the referral to diagnostic investigations.

6.6 Medical information service

Both when searching for a diagnosis and setting up the treatment the physician may have to consult books and other written resources. Finding the right information in a short time can be difficult, and such activities are therefore often reduced to a minimum. An important telemedicine service would be to make available for the physician an electronically stored medical library which is continuously updated.

The simplest scenario of a medical information service includes two roles, a requester of information and a provider of information. The requester requests the provider for some type of information, and the provider returns a report with the information that matches the request. The communication between the requester and the provider should be

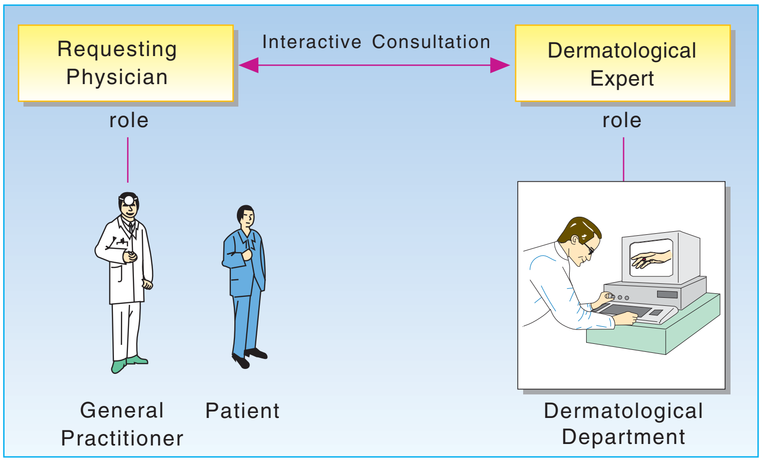


Figure 9 Remote dermatological diagnosis

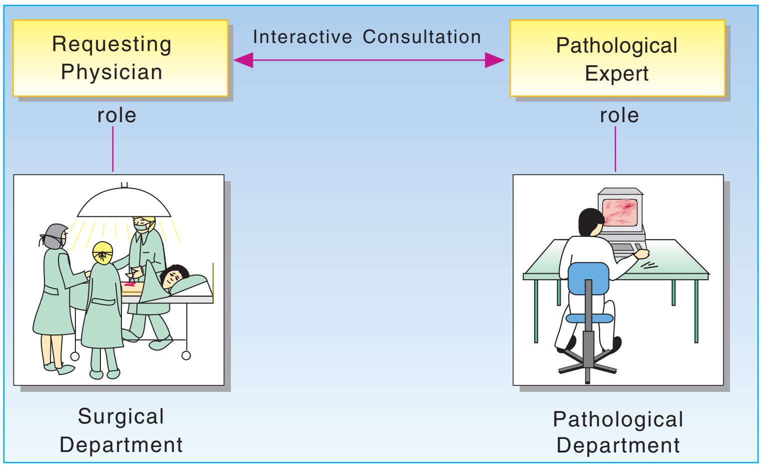


Figure 10 Remote diagnosis of frozen sections

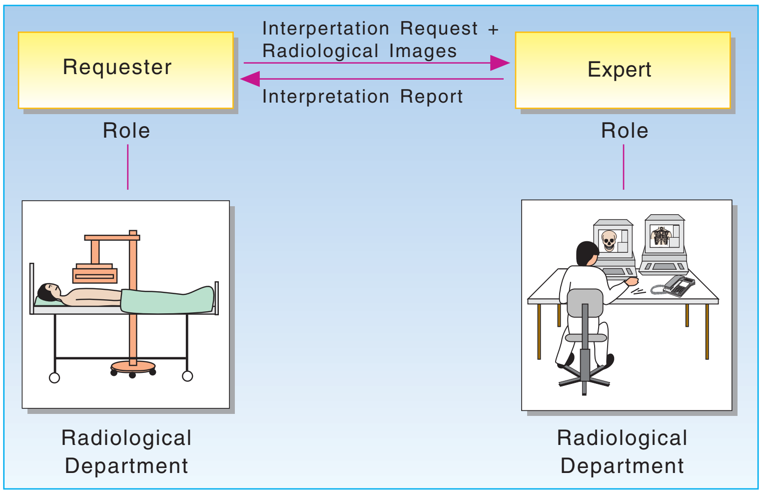


Figure 11 Interpretation of results of radiological investigation

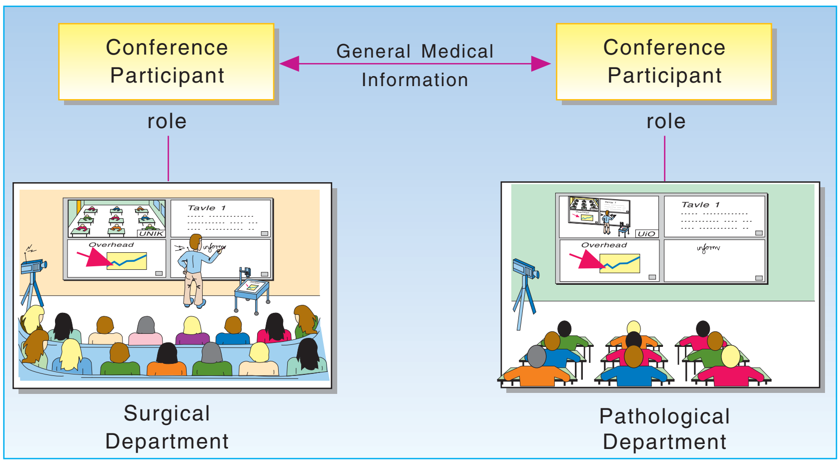


Figure 12 Clinical pathology conference

Table 1 Categories of telemedicine services

	Telemedicine Services	Comm. Roles/ Communicators	Message types/ information types	Characteristics of Communications	TM Objectives
1.	Request of General Medical Services				
	Prescription of drugs	Requesting physician <-> Provider of medicine	Prescription Text information	Non-interactive communication	TM4 TM6
	Example	GP <-> Pharmacy			
2.	Distribution of Patient Information Services				
	Information to Patient distribution	Care provider <-> Patient	Patient Letter Text information	Non-interactive communication	TM4
	Example	GP <-> Patient			
3.	Laboratory Investigation Services				
	Requesting of Clinical Chemistry Investigation	Requester <-> Laboratory Service Provider	Laboratory Service order, Laboratory Service report, Mainly text information	Non-interactive communication. Sample send by ordinary mail	TM3 TM5 TM6
	Example 1	GP <-> Clinical Chemistry Laboratory			
	Example 2	Surgical Dep. <-> Clinical Chemistry Laboratory			
4.	Referral of Patient Services				
	Referral of Patient to Radiological Investigation	Requester <-> Radiological Diagnostic Provider	Radiological Service Order, Radiological Service Report Text Information	Non-interactive communication	TM3 TM5 TM6
	Example 1	GP <-> Radiological Department			
	Example 2	Surgical Dep. <-> Radiological Department			
	Referral of Patient to Hospital	Requesting Physician <-> Specialist	Letter of Referral, Discharge letter/ Mainly text info.	Non-interactive communication	TM3 TM4 TM6
	Example	GP <-> Surgical Department			

	Telemedicine Services	Comm. Roles/ Communicators	Message types/ information types	Characteristics of Communications	TM Objectives
5.	Distribution of Medical Knowledge and Information Services				
	Medical Information Services	Requesting <-> Information Provider	Request for info., Information Report Multimedia info.	Interactive communication	TM7
	Example	GP <-> Medical Library			
6.	Diagnostic/Consultation Services				
	Remote dermatological diagnosic	Requesting Physician <-> Dermatological Expert	Video conference with exchange of skin images. Multimedia info.	Interactive communication	TM1 TM2 TM5
	Example	GP <-> Dermatological Department			
	Remote diagnosic of frozen sections services	Requesting Physician <-> Pathological Expert	Video conference with exchange of images of frozen sections. Multimedia info.	Interactive communication	TM1 TM2 TM5
	Example	Surgical Dep. <-> Pathological Department			
7.	Interpretation of Results of Investigation Services				
	Interpretation of results of radiological investigation	Requesting <-> Radiological Expert	Interpretation Request Radiological images, Interpretation report. Multimedia info.	Non-interactive communication	TM1 TM2 TM5
	Example	Radiological Dep. at logical Hospital <-> Radiological Dep. at regional Hospital			
8.	Remote Education/Conference Services				
	Clinical Pathology conference	Conference participants	Video conference with wxchange of medical knowledge. Multimedia info.	Interactive communication	TM7
	Example	Surgical Dep. <-> Pathological Department			

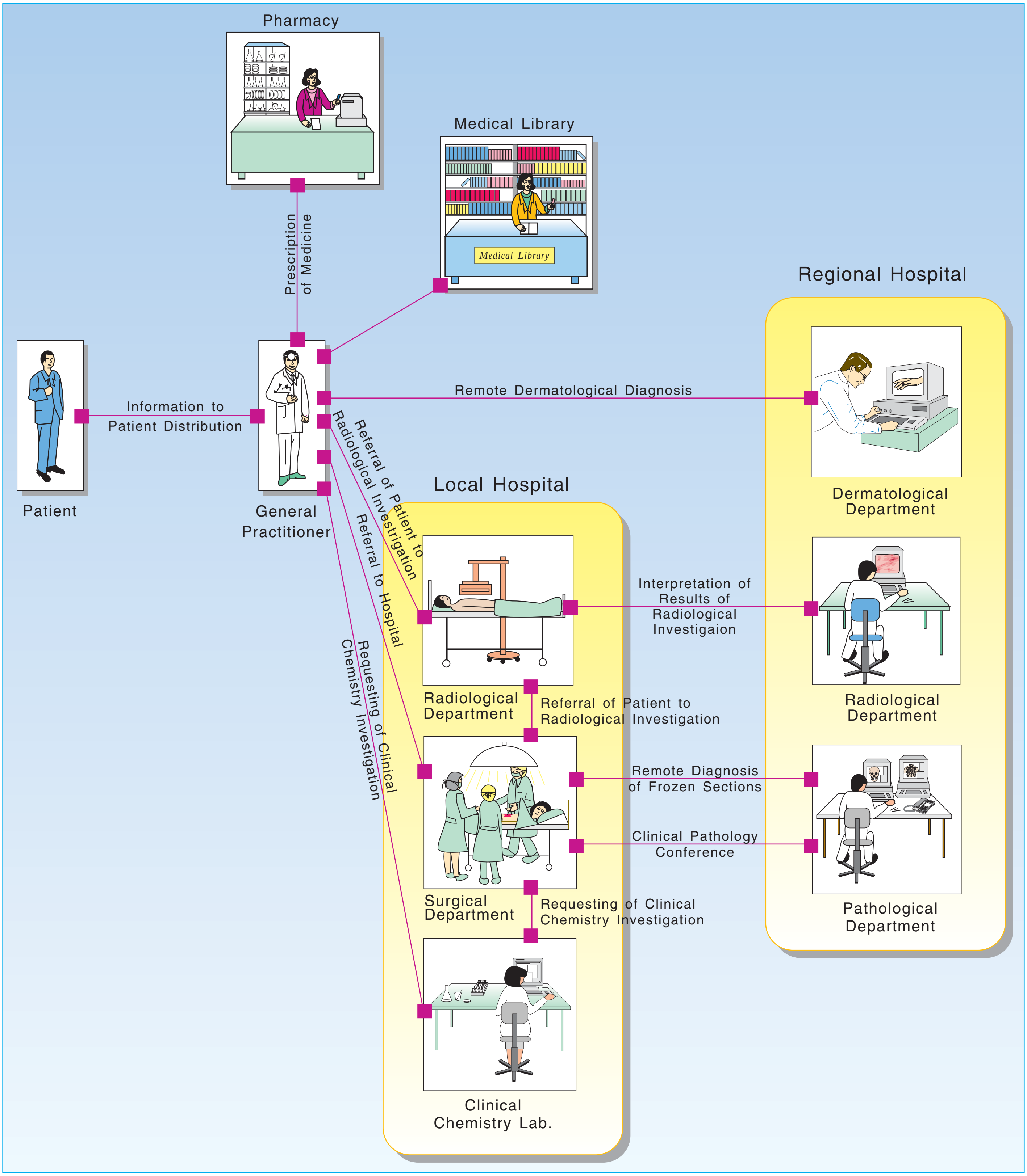


Figure 13 Example of a regional health care network

interactive in order to allow the requester to do repetitive search for information. In our example, the requester is a GP, the provider is a medical library.

This service fulfils the objective of making medical knowledge more accessible (TM7).

6.7 Remote dermatological diagnosis

A physician may need to consult an expert before he makes the patient's diagnosis. Today, experts are contacted through the use of a telephone, but this only allows for solving some of the problems. By using a remote diagnostic/consultation service with interactive video it is possible for a remote expert to participate in the patient consultation. This scenario includes the roles requesting physician and the expert which are interactively exchanging information about the patient. Depending on the situation, the patient may or may not be directly participating in the consultation.

In our example a GP consults an expert at the dermatological department of a hospital. Images of the patient's skin is transferred from the GP to the dermatological expert who diagnoses them and gives his advice for treatment. The GP and the expert are able to discuss the images together. The patient needs to participate in the consultation in order for the expert to select those images he requires.

This scenario fulfils a number of the telemedicine objectives. The benefit of treating a patient as close as possible to his home is met (TM1). The expertise at the dermatological department as well as the pathological department is made available to other parts of the health care sectors (TM2). Experts are shared without high travelling costs, thus resulting in better use of human resources (TM5).

6.8 Remote diagnosis of frozen sections

Another scenario for doing remote diagnosis is the remote diagnosis of frozen sections. The scenario consists of a requesting physician and a pathological expert. In our example a physician at a surgical department of a hospital consults a pathological expert at another hospital. During a cancer operation a sample is taken from the patient. The sample is cut into slices and prepared for microscope investigation. By use of video conferencing and remote microscope steering, the pathological expert can control the microscope and get those images he needs to have a detailed look at. The pathological expert makes a diagnosis and advises the surgeon on what to do next.

The telemedicine objective met by this scenario is the same as for the dermatological scenario (TM1, TM2 and TM5).

6.9 Interpretation of radiological investigation

If the physician is able to perform an investigation but not to interpret the result, he may forward the result to an expert for interpretation. A scenario which describes this situation includes the two roles, requester and expert.

In our example the requester and the expert are both physicians at departments of radiology, but at different hospitals. The requester takes the images of the patient and transfers them to the expert at the other hospital. The radiological expert interprets the images and makes the diagnosis. The diagnosis is returned to the requester in an investigation report. The communication between requester and expert is non-interactive and the request and report are sent in separate sessions.

This type of service reduces the need for the patient to travel to a large hospital in order to perform a radiological investigation (TM1). The images can be taken at the local hospital and exchanged for interpretation by an expert at another hospital. As for the remote diagnosis/consultation services this service makes medical expertise more easily available (TM2) and the ability to share medical experts higher (TM5).

6.10 Clinical pathology conference

In addition to provide care for patients, the health care workers have the responsibility to update themselves continuously. This is today done by medical literature, workshops, and education programs. Participation in such meetings normally requires the health care worker to leave the working place and be out of work for a period of time.

By use of video conferencing, meetings can be arranged without travelling. The cost is presumable lower than traditional travelling and in case of an education program, meetings could be held more rapidly. In this scenario we have one or more parties with the same roles, conference participants. General medical information is exchanged between the conference participants.

Our example is the clinical pathology conference where participants from a surgical department of a local hospital discuss a number of pathological cases with experts form the pathological department of the regional hospital. The aim of the meeting is not to follow up any particular patient but to teach the physicians at the local hospital to interpret pathological images themselves.

Distant education/conference services makes medical knowledge more accessible (TM7).

7 Categories of telemedicine services

Based on the previous analysis of some telemedicine application, a list of categories for main types of telemedicine services may be created. Table 1 identifies eight different categories of telemedicine services.

8 Integration of telemedicine services into a health care network

In the two previous chapters individual telemedicine services and their applications have been described. We started this paper with a short presentation of the regional health care system in Norway. In order to reach the objectives of telemedicine the presented applications must be integrated in a network that supports the activity in the regional health care system.

During the presentation of the various telemedicine applications, we have indicated the parties who are involved. We now take the view of the parties and look for which roles and applications they participate in. Based on the result of this analysis we create a network for the communication in a health care region.

Figure 13 shows an example of a regional health care network. The parties are organised in such a way that primary care is found on the left side and secondary care is situated on the right side.

9 Conclusions

This paper has defined seven objectives for the introduction of telemedicine. It has outlined the structure of the regional health care system of Norway. It then selected three parties in the regional system: a GP, a surgical department of a hospital and a radiological department of a hospital for an analysis of their needs for telemedicine services. Based on their needs, a total of ten different telemedicine applications were analysed and described. These were later categorised into groups of similar characteristics. Finally, the ten applications were put together forming a regional health care network.

The number of health care parties analysed in this paper has been limited for practical reasons. However, even with this restricted selection, the telemedicine applications described fulfil the objectives introduced at the beginning of this paper. This means that our example of a regional network as a whole meets the objectives.

This paper limited its scope to analysing the regional health care system. The need to communicate in the Norwegian health care sector extends beyond communication within one region. A health care network should also cover the need for communicating between regions and with national health care institutions.

In a situation where the possibilities for medical treatment improves much faster than the availability of resources, a new way of organising the health care is required. The concept of a network between the parties in a health care region is rather new and promising. By establishing such networks medical knowledge and resources will be more easily available and can be more efficiently used within the regional health care system.

In this paper we have mainly focused on doing a horizontal analysis of the regional health care system. By horizontal we mean that a number of needs of the parties have been briefly analysed to get an overview of their roles, functions and possible services. Contrary to a horizontal analysis, a vertical analysis would have selected one particular scenario and detailed this in order to find a concrete solution. A horizontal analysis may be followed by a vertical one for particular applications.

A vertical analysis of each application in the network is necessary prior to determination of the detailed technical infrastructure. However, the lists of categories of telemedicine services give some directions. In order to reduce the cost of the infrastructure the technical communication platform should be reused as much as possible. Eight of the twelve applications require only non-interactive communication. These applications may use, as an example, a store-and-forward communication protocol such as X.400 as a common communication platform. The challenge both for the Norwegian health care community and Norwegian Telecom is to establish practical solutions for an infrastructure for telemedicine services based on international standards.

By looking at the information types for each application five out of twelve applications require handling of multi media information exchange. Seven applications require only text-based interchange formats, but all the analysed parties are involved in some application that requires multi-media information exchange.

Horizontal and vertical analysis are important tools for defining which health care and technical standards are required to establish interworking computer systems in the health care. Scenario descriptions already play an important part of the electronic data interchange (EDI) standardisation work both inside and outside of the health care. The scenario description method is likely to be further developed in the coming months.

A major challenge is the requirement to integrate different applications at one party site. Two applications may have different health care and technical requirements. A party that needs to use both applications will have to implement the synergy of the requirements for each application. In our network the GP participates in seven applications. A standardised, yet flexible way of doing this integration is then required.

When defining a strategy for introducing IT and telemedicine applications into health care the network concept described here should be a fundamental basis. By using the horizontal analysis method, the communication needs in the health care can be identified. The result may be used to decide which application types should have the highest priority when building a health care network.

References

(AIM, 1990) Advanced Informatics in Medicine (AIM), Commission of the European Communities. Supplement Application of the Telecommunications for Health Care Telemedicine, Ref AI, 1685, 1990.

Hidas, P. 1992. *HIT-prosjektet: Helsevesen og informasjonsteknologi. versjon 2.* Oslo, Stradec, Sosialdepartementet.

Pankoke-Babatz, U. 1988. *Computer based group communication: the AMIGO Activity Model.* Ellis Horwood. ISBN 0-7458-0694-5.

Report on the Open-EDI Conceptual Model. SIS-ITS Information Technology standardisation. ISO/IEC JTC1/SWG-EDI N222-1, 21 May 1991.

Stenvold, L A. 1992. *Kommunikasjonsstrukturer i helsevesenet. En gjennomgang av kartleggingsforsøk i Telemedisin-prosjektet.* Kjeller, Norwegian Telecom Research. (TF report R7/92.)

Telecommunication for remote consultation and diagnoses

BY GJERMUND HARTVIKSEN AND EIVIND RINDE

621.39:61

1 What is remote consultation?

Remote diagnosing is described by Nymo (1993) as a medical practise where a medical specialist makes a diagnosis based on transmitted information about the patient. Remote consultation is an extension of the concept by the fact that it also includes treatment of patients and furthermore, it is not limited to medical specialists.

The idea of performing medical examinations and evaluations through the telecommunication network is not new. Shortly after the invention of the telephone, attempts were made to transmit heart and lung sounds to a trained expert who could assess the state of the organs. However, poor transmission systems made the attempts a failure (Reiser 1978). Later on, several attempts were made to transmit electrocardiograms through the telephone network (Einthoven 1906). As the distance was only a few kilometres, the practical value of the experiments with remote consultation was not significant.

After World War II techniques for transmission of pictures were developed and hence interest for transmission of medical pictures aroused. This was mainly reflected in a series of individual experiments by pioneers and enthusiasts with medical background (Lovett and Bashur 1979). Among others, attempts were carried out with teaching in different medical topics (Gladston 1954), remote diagnosis and remote consultation within psychiatry (Wittson et al 1961) and radiology (Jutra 1959). These attempts focused strongly on the technical sides of the experiments, but demonstrated at the same time that telecommunication to some extent can replace personnel and patients' need for travelling (Nymo 1991).

Through development of the computer technology during the eighties the horizon for remote consultations was significantly extended. Today, applications for remote consultations depend heavily on digital image processing. Equipment for medical examinations can either output digitised pictures directly, or video signals which can be digitised. Within radiology there are work stations that can be used for traditional X-ray examinations as well as computer tomography and magnetic resonance (Olsson 1989). While the experiments on remote consultations before 1980 mainly were carried out in North America, experiments were performed during the eighties on teleradiology in Japan (Heshiki et al 1988) and Denmark (Wenzel and Frovin 1988), expert consultations within various medical fields in Italy (Cellini et al 1987) and remote diagnoses within paediatrics in India (Desai et al 1984).

2 Why remote consultations?

Why should we invest resources in the development of remote consultations, are not the "near consultations" functioning well enough? If not, what problems can be solved by remote consultations?

The answer to these questions must be found on the basis of some of the challenges met by the public health service. In Norway the responsibility for health services is divided into a hierarchical structure of three levels, viz. institutions for primary health care services, local hospital services and the national, regional and specialist hospital services. The general principle within this system is treatment on the lowest level of efficient care.

This means that the population with need for medical care shall seek contact with the health service through a general practitioner (GP). This doctor will examine the patient. If it is considered necessary to involve specialist competence, the patient is referred to a higher level in the system. The reference is routed to the nearest topical specialist, who may be located at a local, regional or central hospital, or it may be a specialist with private practice.

Specialisation must necessarily lead to a certain extent of centralisation. The more specialised a doctor is, the greater part of the population he has to serve. To maintain a speciality it is required to have a sufficient number of patients. The centralisation entails a drawback for areas with scattered population and with a general lack of specialists, as contact with specialists often involves long waiting time and long travels. This leads to an unequal distribution of health services offered (Kinseth et al 1989).

As an example, Oslo had in 1992 one specialist per 291 inhabitants while the corresponding number for Finnmark was one per 1,194.

To meet these two challenges, viz. the increased specialisation and the health-political desire to have a decentralised and locally adjusted offer of health services, remote consultations can be one of several means. Remote consultations evade the limitations due to geographical conditions as the signification of physical presence is reduced. It is thus possible to reduce waiting time and travel time for the patients. This makes it possible to decentralise parts of the health services as the patient not always must see the specialist face to face. They can meet via telecommunications.

Another benefit inherent in the concept is the possibility of organising specialist pools. As an example, several local hospitals can co-operate in the establishment of a specialist position and by help of remote consultations cover the needs of all the hospitals involved.

A third benefit by implementing remote consultations is to reduce the time factor. Proper equipment and adequate routines increase the possibilities to consult ade-

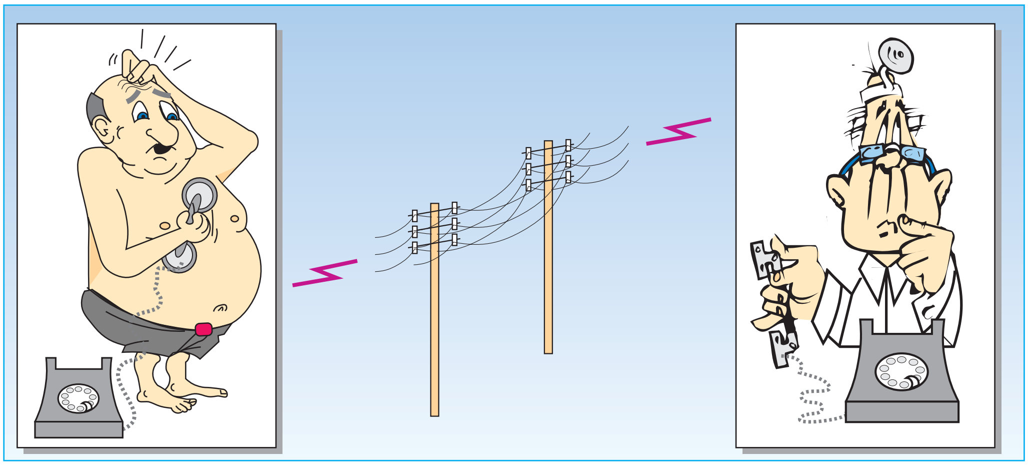


Figure 1 Remote consultation at an early stage

quate expertise in case of immediate help.

The introduction of remote consultations may also entail the possibility of grading the importance of specialists. This is illustrated by the following example. A patient with a sore throat first calls on a general practitioner with his troubles. The GP examines the patient and makes his assessment based on his qualifications. A possible assessment could be that the patient should be examined by an ear-nose-throat specialist. Then the GP sends a reference to a specialist. This twofold reference/ not reference entails some weaknesses. Among other things, the outcome depends strongly on the experience of the GP in question. By help of alternative forms of remote consultations it is possible to introduce a graded specialist influence. The first level covers the cases where the GP feels certain of the treatment he himself can give. The next level is to have a specialist's assessment whether consultation is required. The base for such assessment could be transmitted still pictures.

If the specialist deems it necessary, he goes further to a remote consultation based on video conference.

On the last level the patient has to meet to an ordinary specialist consultation. The effect of such a process is that the patient feels confident about getting an adequate medical treatment. For the GP it means that he is able to offer an adequate solution. Further, it allows a transfer of competence from the specialist to the GP.

3 How to perform remote consultation

Remote consultation can be performed within one of the three levels in the Norwegian public health service, or it can be carried out between two levels. The subject of remote consultation may be direct treatment of a patient, instruction or teaching, where the health personnel are present at both locations.

Telephone inquiries from patient to a doctor are also a kind of remote consultation. With one exception, this paper will not elaborate on this kind of remote consultation.

The way remote consultations are performed varies over a wide range. One extreme is within psychiatry where skilled personnel at a child guidance clinic treat children or whole families through two-way sound and picture connection. The conversations take place with or without skilled personnel present together with the patient. For this process to function properly, the psychologist/-social worker must know the patient she is treating. The conversation itself is most important. It is also an advantage if the remote consultations can be supplemented by "near-consultations".

An example of the other extreme may be teleradiology, where the X-ray is taken and transmitted together with the requisition. In turn, this is assessed by the radiologist to a certain extent independent of information about the patient or in which environment the picture had been taken. Most remote consultations are performed somewhere in between these two extremes, where medical pictures might be supplemented by relevant information about the patient.

When developing methods for remote consultations, one must take as a starting point the topical speciality and see how consultations are performed locally and what sort of equipment is employed.

A lot of medical equipment producing images exists, but not all of it produces images that are adapted for direct teletransmission. X-ray instruments e.g. produce images on analogue film. Applications for remote consultations employed are thus rather diversified. In some specialities it is sufficient with still pictures, while in other cases live pictures are required. In some cases video sequences of examinations can be recorded and transmitted, while other examinations require instruction and guidance by the specialist. The demand for resolution, colours or shades of grey, as well as demand for dynamics of the picture will vary. In addition, there are both analogue and digital picture sources. All these various demands put conditions on the choice of tele-technical solutions.

4 Remote consultations in Northern Norway

During the time NTR has run the telemedicine project, experience has been made through a series of experiments and development work within various medical fields. This includes development of methods and products with the objective of giving smaller health service units access to specialist services, independent of where such competence is located. A main collaborator has been the University Hospital of Tromsø (UHT). This paper basically focuses on experience from remote consultations between regional and local hospitals, or between regional hospitals and the primary health service.

Transmission of images via different kinds of telecommunications, mainly video conferences and data communication, is a central issue in remote consultations. Development has reached varying stages within individual medical disciplines. In some disciplines there are products for remote consultation already in daily use, whereas others are still in an early phase of development.

5 Radiology

In radiology we have experimented with expert consultation by video conference (VC), and with high quality image transfer between workstations. The consultations via VC have not been on a regular basis. The image quality in VC is not good enough for displaying all details and grey levels in i.e. a thorax image. By zooming into the region of interest and by grey scale manipulation, it has been useful, however, for remote education and some kinds of remote consultation where more time than usual may be used on the consultation.

To obtain diagnostic quality for a routine teleradiological service it is necessary to use high quality digitisers and monitors. We have replaced a weekly ambulatory service by radiologists from UHT to a small regional clinic, Troms Military Clinic (TMS). The new procedure involves daily scanning of analogue films, transmission of the digital images to UHT, diagnostic examination on a multi-screen workstation, which is described in a separate article, and digital transmission of dictated reports back to the local clinic (see figure 2). We have designed new work procedures, and installed the necessary equipment.

5.1 The TMS trial

TMS is situated at Setermoen, which is 160 km, or 2.5 - 3 hours car travel one way from Tromsø. X-rays are taken daily, serving both the military and surrounding health centres. The average load is 25 patients per day. In the traditional procedure, a senior radiologist from UHT travels once a week to TMS to describe and diagnose the previous week's 120 examinations. Our project was aimed to completely replace this ambulatory service with the transmission

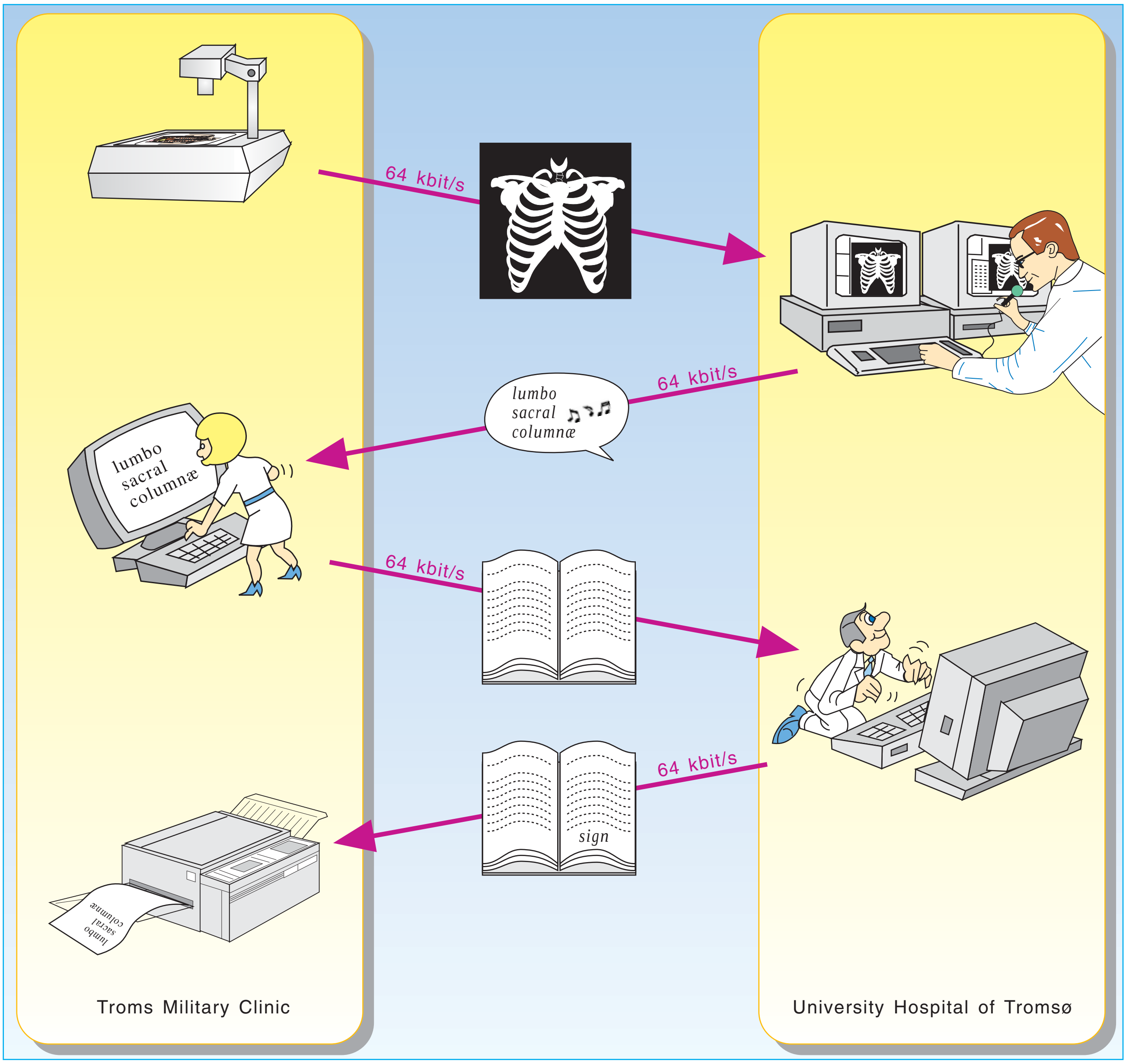


Figure 2 Teleradiological method

of digitised X-ray pictures over 64 kbit/s leased line.

In contrast with most other implementations of teleradiology, the TMS experiment involves a continuous, large flow of pictures, more than 100 per day, and the diagnosis on the computer screen is final, with no viewing of the analogue films afterwards. This places severe requirements on the speed and ease of use of both the digitisation and the viewing equipment. The pictures must be of very high quality. The transmission line must be utilised efficiently, since we operate close to its maximum capacity.

5.2 Digitisation

At TMS a linear scanner, Eikonix 1412 from Kodak Inc., was installed for digitising the analogue films. It was connected to a Macintosh IIx with 8 Mbytes memory and 230 Mbytes disk. We have developed special software for controlling the scanner, so as to read the images fast, with sufficient resolution, and with a minimum of operator intervention.

Each image is read at the scanner's maximum resolution of 4096 x 4096 x 12, but reduced in software from 24 Mbytes to 2 Mbytes by summing squares of 16 12-bit pixels into one 16-bit pixel. The computation is overlaid on image reading, and by programming the lowest levels in assembly the computations manage to keep up so that the scanner is driven without breaks, with time left over for also summing the image histogram.

The image is subsequently transformed to 8 bits depth by histogram equalisation (also programmed in assembly). Having 16 bits image depth allows considerable latitude in image exposure, and it is therefore not necessary for the operator to do any pre-scan exposure measurement. The final image stored on disk is 1024 x 1024 x 8. With more than a hundred images to be scanned each day, each second counts. Our program is much faster than the software that was originally delivered with

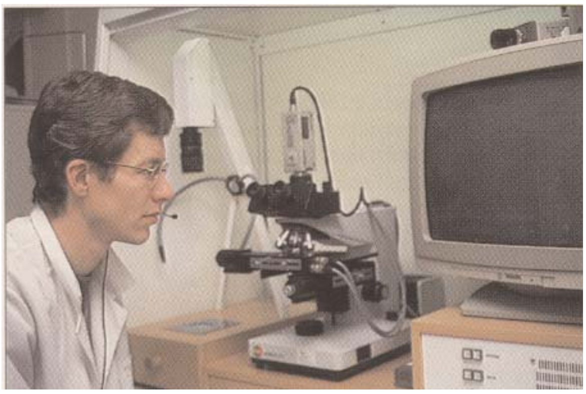


Figure 3 The pathologist working with the pathology workstation

the scanner, but still each picture takes about 90 seconds to scan.

As a special case, the hand-written requisition sheet is also scanned, under ordinary lighting conditions and at much higher speed (less resolution). We have been experimenting with binarisation techniques to turn it into a b/w picture suitable for later reprinting with the written diagnosis overlaid.

5.3 Transmission of pictures

We currently use a 64 kbit/s leased line for the transmission from TMS to UHT. The computers are connected via Ethernet bridges. The bridges and the leased line will be replaced by IP router for ISDN. The transmission uses FTP on TCP/IP, and is performed unattended by the communications software (MacUWS for the Macintosh, Novell Inc.). Each image is transmitted as an ordinary file whose name is composed of patient number and picture number.

The pictures are received by the UNIX-based view-station on UHT. The incoming files are temporarily queued in a special directory, and subsequently reorganised with one separate directory for each patient. Dummy files named only after the patient number act as semaphores between the sending Macintosh and the receiving UNIX system. The part of the project cocerning communications has presented few problems, and the FTP protocol operates very close to the line speed. We have not yet found it necessary to use compression techniques (apart from the data reduction during scanning).

5.4 Training

We started with a training period of two weeks, both for the scanner and for the diagnostic workstation. The personnel found the scanner program easy to understand, and could quickly operate it alone. However, it took some learning to get good pictures routinely. This is due to the need to manually position and orient each film, and manually focus and zoom the scanner objective. The operator should also know enough about an examination to select and scan any relevant old pictures in the patient folder.

Most of the radiologists were pleased with working in front of the computer screens. The program's simple controls made it easy to learn, and the radiologists operated the program independently after having trained 1 - 2 sessions of 15 minutes each.

5.5 Diagnostic quality

When the analogue and digital pictures are put side-by-side for comparison, one can normally see on the computer screen all details from the film. However, there is a certain lack of "depth" in the digital pictures, so that it takes some manipulation with the grey-scale window to see all details. Working in a darkened room is imperative. The possibility to manipulate contrast was used extensively by some of the doctors, thereby enhancing details that might otherwise go unnoticed. Picture inversion was also used, especially on lungs.

During a period of 6 weeks, pictures from 5 patients were scanned and transmitted daily. These were also diagnosed by the ordinary film routine at TMS. Different doctors viewed the analogue and the digital pictures. Preliminary analysis of the results show no significant difference between the two diagnoses. The variations that are found are clearly attributable to differences between the observers.

5.6 Production test

For the scanning, sustained operator speed was about 3 minutes/picture, 4 hours each day. Because of the need to manually adjust picture and camera, it is difficult to do anything else between scans. We will later experiment with reducing the number of scan lines to 2048 or even to 1024 for a corresponding increase in scan speed. We also timed the viewing of the pictures by the radiologists. There were large variations between the doctors, but each doctor worked almost as fast in front of the screen as with films. Grey-scale windowing took some extra time, but this was compensated for by not having to manipulate the physical films.

Today, the weekly ambulatory service has been totally replaced by the teleradiology service. Since October 1992 there has not been any radiologist visiting TMS for diagnosing X-rays. Up to March 1993 X-rays from 2,200 patients have been viewed on the teleradiology system only.

The TMS project shows that teleradiology can now be implemented routinely for serving local clinics, with little or no loss in picture quality, and substantial gains in care quality.

However, replacing a well established procedure with new techniques involves not only equipment, but people and organisations as well. In this description we have concentrated on the technical problems and their solutions. Taking care of the other two aspects is equally important for the successful realisation of our goal: Teleradiology for better health care.

Results so far show that teleradiology may represent a viable solution for small clinics lacking qualified radiologists, as well as clinics with a single radiologist who needs ready access to colleagues. The project also shows that teleradiology can now be implemented routinely for serving local clinics, with little loss in picture quality, and substantial gains in care quality (Sund et al 1991a), (Sund et al 1991b), (Sund et al 1991c).

6 Pathology

Pathology is the medical study of disease-related changes in cells and tissue. Telepathology is defined as the practice of pathology at a distance. The pathologist sees images of tissue on a monitor rather than viewing the tissue specimen directly under a microscope. The concept of using video microscopy to provide pathology services to remote hospitals was first experimented with about twenty years ago (Weinstein 1991). It is only during the last couple of years, however, that telepathology systems have been used regularly in a few hospitals.

The pathological department at UHT has established remote frozen section service to two local hospitals in North Norway (Nordrum et al 1991). The local hospitals are furnished with a VC-connected[1)] workstation equipped with a motorised robotic video-microscope, figure 4. The microscope is controlled from the pathological department at UHT, and the video signal from the microscope is transferred by the VC-system. The pathological department at UHT continuously receives video images (dynamic images) and, on demand, still images on the monitors. This system is to our knowledge, the first static-dynamic robotic telepathology image system put into practice.

With this system we can provide small rural hospitals with diagnostic services as if a pathologist were on location. The system is primarily used to provide hospitals with immediate tissue diagnosis on patients undergoing surgery. Laboratory technicians at the local hospitals have been trained to prepare removed tissue for microscopy according to standard frozen section procedures. With limited training the laboratory technicians have been able to prepare frozen sections of good quality.

The validity of frozen section diagnosis based on video microscopy has previously been tested in the pathological department at UHT on archived material. The results of this, and another study, indicate that the accuracy is acceptable for a frozen section service based on video-microscopy (Eide et al 1993), (Weinstein 1991), (Weinstein et al 1987).

The workstation at the university hospital (basically a VC unit), also has a motorised robotic microscope. With this equipment the university hospital offers clinical pathologic conferences to hospitals with VC facilities. Our telepathology system also supports on-line consultations with other departments of pathology having compatible equipment.

The frozen section diagnosis is a preliminary diagnosis. The removed tissue is also sent to the pathology laboratory for a final diagnosis based on a more accurate and time consuming technique. The difference in the diagnoses is shown in table 1 for the 68 frozen section diagnoses made so far. The difference between frozen section diagnosis and final diagnosis in telepathology does not differ significantly from ordinary frozen section service and final diagnosis.

The only neuro-pathologist at the UHT has been a visiting scientist at the Mayo clinic, Rockester US, this past year. In order to enable him to continue providing some of his service to UHT, we have transferred digitised images from the electron microscope at UHT to Mayo. As a back-up and control he also receives the usual paper print by ordinary mail. The results from this trial are promising, and the techniques may be used in expert-to-expert consultations.

Table 1 Correlation between Frozen section and Final diagnosing

		Final diagnoses			
		Malignant	Benign	Deferred	**Total**
Frozen	Malignant	12	1	0	13
section	Benign	2	50	0	52
diagnoses	Deferred	1	2	0	3
	Total	15	53	0	68

7 Otorhinolaryngology

Examinations of ear, nose and throat (ENT) by use of endoscopy is gradually replacing other methods. The equipment utilised to make endoscopic examinations consists of the light source, endoscopes, camera, camera controller and a monitor. In remote consultation, this equipment is integrated with a video conference unit. A general practitioner with a short introductory course in endoscopy controls the endoscope. He brings his patients to the local studio and can from there communicate over the two-way audio and video connection with the specialist located in the studio at UHT. The specialist can see the endoscopic examination on a monitor and influence the control and movement of the endoscope with the GP operating it (see figure 4).

In our first trials, a total number of 17 randomly selected ENT patients were examined. First, they were all examined by simulated remote examination and afterwards, they were examined by ordinary endoscopy. This comparative test showed a complete correspondence between the findings that were made during ordinary endoscopic examination and during a simulated telecommunication

[1)] *A 2 Mbit/s connection. We also experiment with lower speed, e.g. 384 kbit/s.*

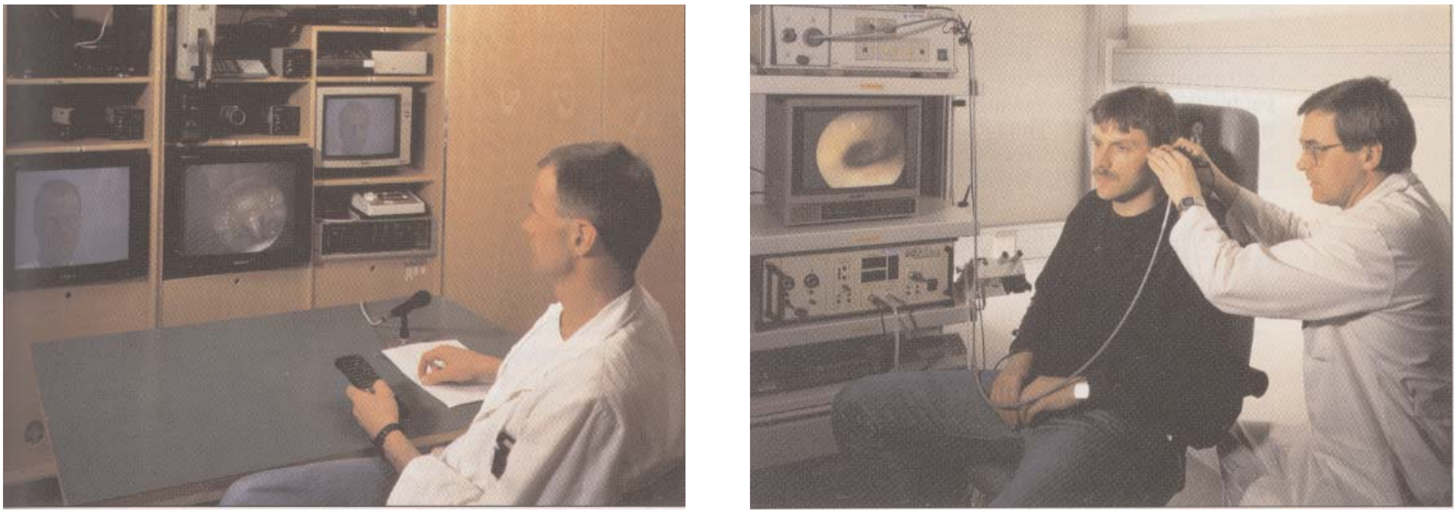


Figure 4 The ENT specialist conducts the examination over the VC network

Figure 5 The cardiologist guides the GP in cardiac examinations

transmission (Hartviksen and Pedersen 1992). After these first trials, approx. 30 patients are examined remotely between studios 240 kilometres apart.

The sound and picture quality are both satisfactory. A remote endoscopic picture containing immobile structures is very detailed. Thus it is possible, for example, to detect small blood vessels along the leg of the hammer in the ear drum. When the remote endoscopic picture contains movable structures, e.g. vocal cords during phonation, the image is accurate enough to exclude/confirm ulcerations and tumours.

The results from the trials so far show that it is quite possible to examine the patient at one site while the evaluation is made on a different site (Pedersen and Hartviksen 1993). So far, it seems like most of the ENT patients applied for specialist consultation, are well suited for remote examination. The ENT specialist did not report great problems with not having the opportunity to lay his hands on the patients, but in the beginning he felt that the time devoted to each patient was too long. This changed gradually and now a remote examination does not last longer than an ordinary endoscopic examination.

The conclusion made by the doctors involved in the trials is that remote endoscopic consultation could have a profoundly positive impact on health care cost, access and quality. What is lacking is not the technical solution, but rather the vision and momentum to apply it on a broad scale.

8 Echo cardiography

Examinations with ultrasonic equipment are important in a variety of cardiac diseases. The examinations are done by specialists in cardiology at UHT. These specialists also provide ambulant services to smaller hospitals in the region.

In remote consultation trials the small hospital is furnished with ultrasonic equipment and VC studio. The physician at the hospital controls the ultrasonic equipment and the position of the transducer. A cardiologist located in VC studio at UHT receives the transmitted video pictures, and make the diagnosis (see figure 5). The cardiologist may guide the examination, and give the local physician advice through the two-way sound and picture connection. Control tests comparing 40 diagnoses made during direct examination and those made via video conferences showed no major differences which could be related to reduced picture quality. The only possible exception may be quantitative measurements of the heart muscle. However, this may be solved by high quality still pictures. Any other deviations are ascribed to the examiner's lacking skills (Afseth and Lunde 1993).

As basic training within echo cardiography this method is well suited. A testimony to the method's success is that it has now been accredited as part of the training program necessary in specialist education. During remote diagnostics within echo cardiography with an inexperienced examiner, the clinical problem can still in most cases be verified or dismissed.

9 Gastroscopy

As in otorhinolaryngology the physician in gastroenterology uses different kinds of video camera scopes to get a view from inside the body. He guides the scope down the throat into the stomach or in rectum, the large and small intestinal. The chip in the scope delivers an ordinary video signal which may be input to video recorders, video conferences, or captured in a computer.

In our project we capture still images with a PC based still image system explained in a separate frame. The images with describing text may be sent to an expert for a second opinion. The other expert could be located inside the hospital or at a remote site. In the future physician will include the images in the patient's medical record for documentation purpose. A reference database built on still images is also an interesting idea in this project.

In addition to capturing stills, video conferences have been used for discussing consultation with colleagues. In the video conference sessions video tapes from the gastro consultations are played. The experiments have not been completed, and more work will be carried out with the use of both video conferences and still images.

10 Dermatology

The dermatologists in North Norway provide an ambulant service to small hospitals and health centres. In a trial, which started in 1989, this ambulant service is replaced by video conferences (Jøsendal et al 1991). A GP and a dermatologist are involved. Twice a month the GP brings his patients to the VC studio. The dermatologist is located in the VC studio at the UHT. The patient tells the specialist his story describing the disease, supplemented by the GP who has previously examined the patient. The camera is focused on the area of skin in question, while the dermatologist views either a live image, or a high quality still image. In the first phase the telediagnoses were "checked up" by a local dermatologist and the two diagnoses were in 100 % agreement. Thus the dermatologists are confident in the diagnoses made through teleconsultation, which up to now have reached the number of 200 (Holand and Stenvold 1993).

Although the experience of the dermatologists have been overall positive, there have been feelings that "something is missing" in this kind of examination. Even though the patient may be satisfied with the consultation, the specialist may feel the loss of some important social aspects between patient and doctor. The specialist may feel there is no natural conclusion to the consultation.

The diagnosis procedure is done collaboratively between the GP and the specialist. In addition to making a diagnosis and proposing a treatment for the patient, the dermatologist transfers some of his knowledge to the GP. In this way the GP increases his level of knowledge through use of telediagnosis. The service is now in ordinary use between the local hospital in Kirkenes and UHT, and will be expanded also to cover lap epicutal testing and ultra-violet treatment over VC studio.

11 Psychiatry

Telebased consultations in the field of psychiatry have been tried in several contexts. Some attempts have not been completed, partly because of poor technical quality of the equipment applied and because central personnel in the experiments have been moved or assigned to other tasks. This illustrates that remote consultations still is a new field, and that the application depends on the interest of individual participants. An example of how remote consultations can be exploited is experience from a successful project within youth psychiatry. In this case the consultation took place between a VC-studio in a psychiatric institution and a studio at the nearest child guidance clinic (BUP) 150 kilometres away. At the BUP a social worker was present in the studio. At the other end were the children, sometimes alone and sometimes together with parents and a local therapist. The therapy consisted of conversations about how the family had solved problems related to the child, and what mother or father had done when they had not understood the child's behaviour or understood how to help it to solve its problems. The great advantage with this remote consultation was considered to be that the therapist and patient were spared from travelling 150 kilometres in order to accomplish a treatment appointment. An additional advantage was that professional instructions were received by the personnel at the local institution. Seven children were treated during the year 1992 and altogether 80 consultations took place.

All parties, the social worker, the local therapists, the children and their families, reported favourably to this form of treatment and instructions (Wilhelmsen 1993). Picture and sound quality were satisfactory and the distance did not critically reduce the contact between the social worker and the patient. It turned out that such a form of encounter was well suited for establishing a dialogue, which is not always the case in a face-to-face consultation. This differs somewhat from what is generally expected from a video conference. An actual meeting between patient and therapist is, by people who have experimented with remote consultations, regarded as the optimal solution. This may stem from the fact that psychiatric treatment is quite different from the cases where somatic treatment has been attempted by remote consultations. Among other things, teleconsultations introduce a state where status and roles are reversed, because the therapist may be as uncertain as the patient as to how therapy through remote consultation actually works. Furthermore, it seems that the video conference situation tends to focus the attention of the participants. It is of no use speaking all at once, and in order to be seen by the other side one has to be in camera focus. Also, it is possible to be let off just by walking away if the situation should become too threatening. For this reason there have been indications that children feel freer than is the case in ordinary therapeutic conversations.

The findings from these experiments show that some problems expected in advance do not come true. Nearness between therapist and patient is necessary but may in some connections become too threatening. Video conferences seem to help in regulating this nearness and may thus be a therapeutic presedence.

12 Microbiology

Bacteria determination is today carried out in microbiological laboratories by sowing specimens on various gel media. After approximately 24 hours bacteria colonies have grown which in some cases can cause changes in the gel media. When the microbiologist assesses the vessels, a typical act is to turn them around to let the light fall from different angles. Little by little he associates appearance and smell of the colonies to a certain type of bacteria. Our idea is that these gel vessels could be photographed for transmission to a specialist within microbiology by a still image system. Control of the results from the experiment under the conditions at that time showed a misjudgement up to 25 %. The reason for the high uncertainty was among others the poor quality of the pictures when we tried to describe three-dimensional patterns in only two dimensions. The colonies' dryness, arching and skin was not picked up by the system. Work is continuing to improve the method. This involves improvement of the camera part and the light set-up. At the same time the pictures should be supplemented e.g. by information of smell.

13 Further challenges

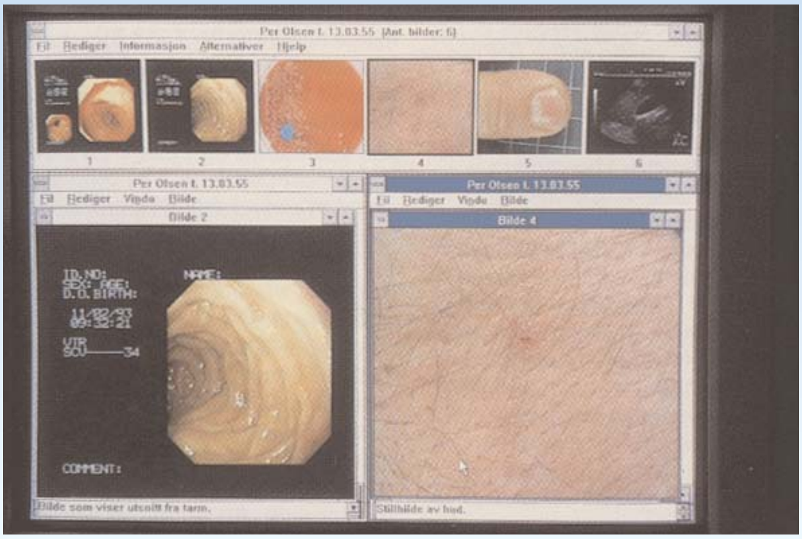
Remote consultations are in the process of being put into regular use. Even if the technical set-ups for remote consultations seem similar, the methods have great variations in their distinctive characteristics. An illustrative limitation because of such peculiarities is the experiment within microbiology.

The greatest challenge in time to come is perhaps still not of technical, but of organising and economic nature. Remote consultations require revision of local routines. Priorities must be changed and legislation adapted. Regulations for economic settlement must to a greater degree take into account the benefits contributed by remote consultations to social economics.

In broad outline, remote consultations have as its objective to move information and/or competence. The public health services are generally concerned with protection of privacy, especially when it comes to introduction of new technology. Sensitive information which can be associated to a patient's identity must not fall into the hands of unauthorised persons. Another factor concerning patient security is a guarantee against mix ups. Concerning transmission of medical information through remote consultations, protection of privacy is fairly well taken care of. On the other hand, it is a wish by the medical profession that information should be stored for later consultations, statistics, research, etc. This introduces a new approach to the problem regarding how this information should be managed. The problem of data storage within the health care is far from solved and the process of archiving images electronically has just begun. Use of remote consultations will be a driving force in this process.

VIDA – a still picture system

The project has developed a PC-based system for the collection, storage, and transmission of still pictures through the telecommunication network. The system is developed especially with a view to being applied within medicine. A standard frame-grabber card and self-developed software constitutes the system. The software consists of three modules which reflect the usage. From a video input source the frame-grabber module continuously grabs the still pictures from the consultation and stores the digitised pictures on the hard disk. Only this module depends on the special hardware. Through the editing module the doctor can inspect the pictures, comment on each picture and on the whole series in the form of text. The system is adapted to transmission by both V.32 and V.32-bis modems and through ISDN. In order to reduce transmission and storage costs series of pictures are compressed. To this end a program based on the JPEG algorithm is employed. The degree of compression varies from 1:30 to 1:50. This implies that a full screen picture is compressed from approximately 1 Mbyte to approximately 20 kbyte. The recipient receives the consultation and can assess the pictures and text associated with the series with the help of the editing module. Among other things, the system is used for experiments within gastroscopy, endoscopy, and microbiology.



14 Concluding remarks

Experiments in Northern Norway over a period of four years have shown that remote consultations are possible. Within some medical fields this method of practising is also in daily use. Remote consultations are still at an early stage, and the examples quoted in this paper are based on just a few experiments. Furthermore, we have reasons to believe that the persons who have been willing to try out remote consultations have been well motivated. Therefore we must be open for the emergence of new applications and new objections when remote consultations are taken into use within other medical fields.

15 References

Afseth, J E, Lunde, P. 1993. *Teleechocardiography: remote instruction in echocardiography via video conferences.* Oral presentation on First International Conference on Medical Aspects of Telemedicine, Tromsø, Norway, 21 May 1993.

Cellini, N et al. 1987. Teleconsultation: four years of experience. *Rays,* 12, 89-95.

Desai, A B et al. 1984. Telemedicine: a feasibility experiment for national application. *Indian Paediatrics,* 21, 773-775.

Eide, T J, Nordrum, I, Stalsberg, H. 1993. The validity of frozen section diagnosis based on video microscopy. *Zentralblatt für Pathologie,* 138, 405-407.

Einthoven, W. 1906. Het telecardiogram. *Nederlands Tijdschrift voor Geeneskunde,* 50, 1517-1547.

Gammon, D. 1991. *Telemedicine in North Norway.* Kjeller, Norwegian Telecom Research. (NTR Report R27/91.)

Gladston, J. 1954. What have radio and television to contribute to continuing medical education? *Journal of Medical Education,* 29, 28-30.

Hartviksen, G, Pedersen, S. 1992. Remote endoscopy of otorhinolaryngology. *SPIE Proc.,* 1649.

Heshiki, A et al. 1988. Teleradiology at Gunma University Hospital. *Medical Informatics,* 13, 295-301.

Holand, U, Stenvold, S E. 1993. *Are the patients satisfied with teledermatological consultations?* Oral presentation on First International Conference on Medical Aspects of Telemedicine, Tromsø, Norway, 21 May 1993.

Jutra, A. 1959. Teleroentgen diagnosis by means of videotape recording (Editorial). *American Journal of Roentgenology,* 82, 1099-1102.

Jøsendal, O et al. 1991. Long distance diagnosis of skin diseases. *Journal of the Norwegian Medical Association,* 111, 20-22.

Kindseth, O et al. 1989. *Sammenligningsdata for somatiske sykehus.* Trondheim: Norsk Institutt for Sykehusforskning, NIS-report No. 7.

Lovett, J E, Bashur, R L. 1979. Telemedicine in USA: an overview. *Telecommunication Policy,* 3, 3-14.

Nordrum, I et al. 1991. Remote frozen section service: a telepathology project in Northern Norway. *Human Pathology,* 22, 514-518.

The telepathology workstation – TELEMED A200

The frozen section procedure is made of different parts.

1 The surgeon cuts a specimen from a cancer suspected organ.

2 The specimen is brought to the pathologist.

3 The pathologist looks at and touches the specimen.

4 The pathologist decides where to make the slice.

5 The laboratory technicians prepare the removed tissue for microscopy according to standard frozen section procedure.

6 The pathologist looks directly at the specimen.

7 The pathologist scans the specimen by use of microscope with a small magnification. He searches the suspected areas with higher magnification.

8 The pathologist makes a preliminary diagnosis.

9 The pathologist informs the surgeon about the diagnosis.

10 The remaining specimen from the cancer suspected organ is prepared and a final diagnosis is made. This preparation takes several days.

The remote frozen section procedure is made as close as possible to the ordinary frozen section procedure. The pathology workstation consists of elements that takes care of the different parts of the procedure.

1 The surgeon cuts a specimen from a cancer suspected organ.

2 The specimen from the organ is put under a macro camera and the surgeon describes what the consistency of the specimen is like. The pathologist turns on and adjusts the over light.

3 The pathologist then instructs the surgeon where to make a slice for the specimen. The pathologist and the surgeon may point at the video image with each others' cursors. The cursors are shown at both places simultaneously.

4 The laboratory technicians at the remote hospital prepare the removed tissue for microscopy according to standard frozen section procedure.

5 The specimen is placed in a special holder. The pathologist turns on the under light and looks at the specimen. A snapshot is taken of the whole specimen and placed in the lower right hand corner of the monitor and will act as a map later.

6 The specimen is then placed under the microscope and the pathologist scans the specimen by use of microscope with a small magnification. He searches the suspected areas with higher magnification. He controls the microscope by use of a kind of big button. The X and Y movement of the specimen stage is controlled by pressing the button left/right and forward/backward. Focusing is done by turning the button left/right. All these controlling movements have two steps for two different speeds. On the big button there are three switches for controlling the illumination, changing the lens (magnification) and for controlling an optical zoom. While moving the specimen stage a mark in the "map" in the lower right hand corner of the monitor shows the actual position. The pathologist may also mark suspected position while he is scanning the specimen. The location of the marks is shown on the "map" and he may go back to this position by clicking, pointing and clicking the mouse on the marks. Images from interesting areas may be saved on the computer's hard disk and shown as small images in the right half of the monitor.

7 The pathologist makes a preliminary diagnosis.

8 The pathologist informs the surgeon about the diagnosis.

9 The rest of the specimen from the cancer suspected organ is sent to the central hospital.

10 The specimen is here prepared and a final diagnosis is made.

The pathologist's workstation is in fact a specialised video conference equipment. In addition to two-way audio and video, which is part of all video conference communication, the equipment must establish a data connection for controlling the remote video switch, microscope controls, the light for the macro camera and displaying of remote cursor at both locations.

Almost every video codec has a data communication interface, and the equipment has been used on video codecs from different manufacturers and at different communication speeds. In addition to video codecs, ordinary video conference equipment consists of video/audio sources, loudspeakers, monitors, and a video switch. In the pathologist's workstation all this is built around an ordinary personal computer and some specialised hardware. The PC acts as the user interface and so changing the user interface for other video conference applications is fairly simple. With a video input card in the PC two relating video images are displayed on the PC monitor simultaneously. These images may be scaled to an arbitrary size.

When the program is started the first screen brings up a list of actual communication partners to select from. When selecting a partner a command is sent through a serial port of the PC to the data terminal equipment, i.e. MegaNet, to set up the connection. When the connection is established the PC will act nearly as an ordinary viewphone. The incoming image is shown as a large image and the outgoing image is shown as a small image in the lower right hand corner. Under the incoming image there are soft buttons for selecting modes.

1 Videophone. In the videophone mode both communication partners see each other face to face.

2 Macro. In macro mode the big control button controls the illumination for the macro camera. The soft buttons make it possible to select over or under illumination, capture a reference image which should be used as a map for the microscope modes and has a button for going to the other modes.

3 Microscope. In microscope mode the big control button controls the movement and adjustments of the remote microscope. The soft buttons make it possible to capture still pictures, mark positions, re-select position, and go to the other modes.

Both parties' monitors display a large image from the microscope, a map of the specimen in the lower right hand corner and any captured small image from the microscope on the right edge of the screen.

The workstation is commercialised by NovaKom A/S.

Nymo, B J, Engum, B. 1990. *Telemedicine to improve the quality, availability and effectiveness of the health service in rural regions.* Kjeller, Norwegian Telecom Research. (NTR Lecture F10/90.)

Nymo, B J. 1993. Telemedicine. In: *Telektronikk,* 1, 4-11.

Nymo, B J. 1991. Helsevesen. In: *Nye nettverk: telekommunikasjon i samfunnet.* Ed: Mathisen, K O, Rasmussen, T, Tetzchner, S v. Oslo, Gyldendal.

Olsson, S. 1989. Ny vår for teleradiologien. *Medicinsk Teknik,* 6, 22-25.

Pedersen, S, Hartviksen, G. 1993. Teleconsultation of otorhinolaryngology patients. *Archives of Otolaryngology - Head and Neck Surgery,* (in press).

Reiser, S A. 1978. *Medicine and the reign of technology.* London, Cambridge University Press.

Sund, T, Rinde, E, Størmer, J. 1991. Full-scale replacement of a visiting radiologist service with teleradiology. *Proc. CAR 91.*

Sund, T et al. 1991. Techniques for multimedia teleradiology. *Proc. CAR 91.*

Sund, T, Rinde, E, Størmer, J. 1991. *The TMS teleradiology experiment.* Kjeller, Norwegian Telecom Research. (NTR Lecture F12/91.)

Weinstein, R S. 1991. Telepathology comes of age in Norway. *Human Pathology,* 22, 511-513.

Weinstein, R S, Bloom, K J, Rozek, S. 1987. Telepathology and the networking of pathology diagnostic services. *Arch. Pathol. Lab. Med.,* 111, 646-652.

Wenzel, A, Frovin, T. 1988. Teletransmission of radiographs: a review and a description of a prototype system. *Tandlægebladet,* 92, 502-509.

Wilhelmsen, S. 1993. *Fjernt, men likevel så nært.* Thesis in social studies. University of Trondheim, spring 1993 (in prep.).

Wittson, C L, Afflect, D, Craig, D. 1961. The family health worker. *Journal of American Psychiatric Association,* 117, 21-39.

Mira – teleradiology and digital radiology

BY TORBJØRN SUND

621.39:61

1 Introduction

Images have always played an important role in medical diagnosis. Imaging techniques that look "through" the body generally belong to the field of radiology, such as X-ray imaging, ultrasound, magnetic resonance, etc. Some imaging techniques have a signal resolution and noise comparable to standard TV. They can be transmitted without significant loss of quality over video conference lines, for teleradiology and primary diagnosis. Our own experience with such images includes ultrasound images of the heart, and low-resolution CT-images.

X-ray films however, have much higher resolution than television. As an example, films used in mammography can resolve more than 20 line pairs per millimetre (lppm). A 15 x 20 cm film displayed on a standard TV would, on the other hand, have a resolution of only about 1 lppm. Also the dynamic range (S/N ratio) of a film is 10-12 bits, much higher than the 6-8 bits found in standard TV. Lastly, when an image is transmitted over a video conference line, it is compressed, and loses in that process much of its high-resolution content. This may be only beneficial for a wrinkled human face, but is unacceptable for an X-ray image.

For teleradiology to be useful in routine diagnosing of X-ray images, one must therefore digitise the images with very high resolution, and send them as purely digital files with no transmission loss. If the sender and the receiver do not need to interact with each other, an off-line, still-image system can be used. If they need to communicate while looking at the images, an on-line, interactive system is needed.

When we started our work in teleradiology, no system was available that could cover these needs adequately. We therefore developed software to digitise X-ray images with sufficient speed and high resolution, send them unattended over a network, view the images for diagnosis on digital computer screens, and communicate between viewers using both images and sound. We call this collection of programs "Mira" (from Latin mirare, view attentively).

In the following the Unix-based viewing station is described. It can be composed of many screens, has integrated sound, and communicates effectively over any network. After our development begun, commercial implementations of still-image transmission systems have appeared. Their emphasis is on the transmission of a few images at a time, using colour TV frame grabbers for digitising, and not very useful for X-ray images. No match for our viewing station has yet appeared. Figure 1 shows a radiologist using a 4-screen Mira system.

2 A multi-screen workstation

Radiologists view many images at a time. Films are hung side by side on large view panels ("light boxes") for comparison: Left/right, old/new, adjacent CT slices, different X-ray projections. A teleradiology system must be able to compete effectively and economically with this. We decided from the start that the images would be displayed digitally, and we have devised a menu system and a terminology that makes it easy to install and work with an almost unlimited number of screens. As a contrasting example, in a teleradiology system in installation in South Korea between small army hospitals and Seoul, the received digital images are reprinted on films before viewing (1). In this system, user acceptance is less painful since the radiologists see films as before, but cost is higher and extendibility less.

Central to Mira's use of multiple screens is the concept of a screen "map". Part of the primary screen (console) is used for projecting miniaturised reproductions of each of the screens that are connected. The content of each map reflects the con-

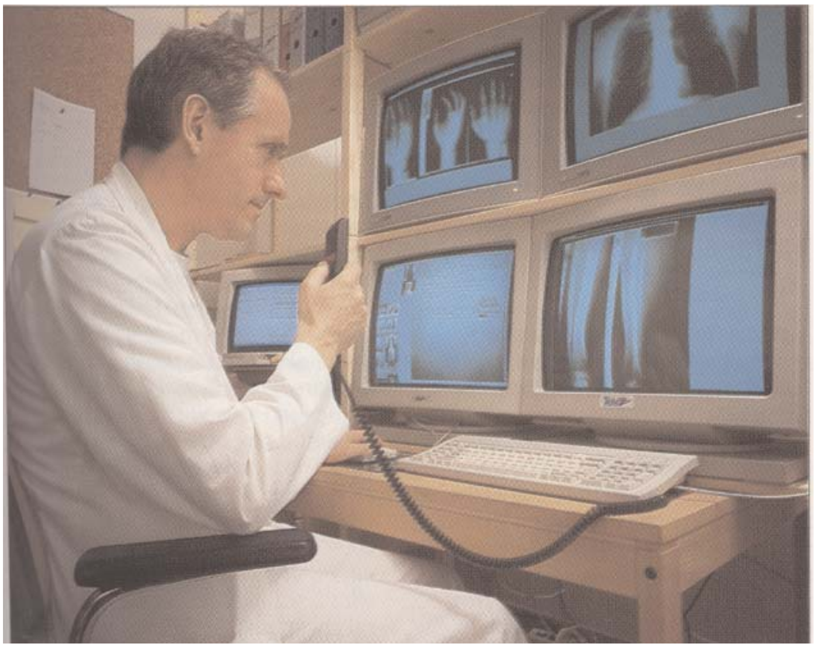


Figure 1 Radiologist making diagnosis from digitised X-rays, viewed on a Mira installation with four screens. All the equipment is off-the shelf. The two lower screens are connected directly to the computer, the two upper screens are networked X-terminals

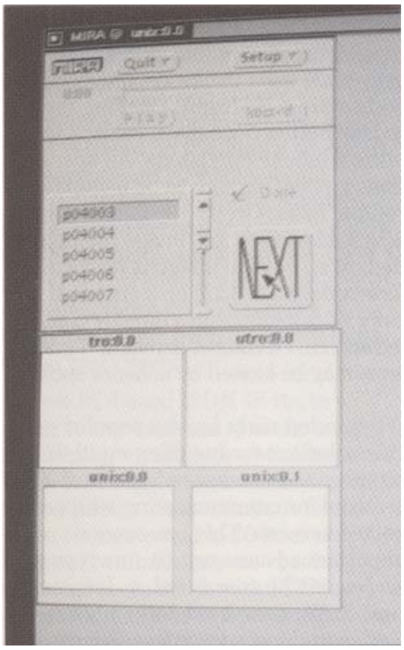


Figure 2 Close-up of console screen in a 4-screen Mira installation. Part of the console screen area is used to show "maps" of itself and the other screens

tent of the corresponding physical screen. The position of the maps is easily adjusted, and will normally be chosen to reflect the relative positions of the real screens. The mouse pointer is used to identify any image in the map, for pointing within and for manipulation. Figure 2 shows an example console screen, with four physical screens connected.

The idea of a "virtual screen", where the paintable area is much larger than the physical screen has been around some time. The concept of screen maps, on the other hand, makes it possible to compress many screens into a small area. Although seemingly opposite, these ideas have much in common, and we owe much of our thinking about the operations in a screen map to "Solbourne Window Manager", the first virtual window manager for X (2).

The operations that are implemented on an image have been chosen according to the needs of the radiologist: Turn and mirror, move, enlarge, dynamic zoom, and grey scale adjustment. Of these, the grey scale window and level is the most used. It is performed by pointing at an image and clicking a mouse button. In this mode, mouse travel is translated into adjustment of grey-scale level (sideways movement) and contrast (to-from). Clicking the mouse button once more returns to normal mouse operations. All operations are fast enough to be useful on a routine basis: Grey-scale enhancement is instantaneous, and image manipulation on a 1 MByte image is performed in less than a second on a standard Unix workstation with a directly connected screen. Of course, networked systems may be limited by network speed.

We decided early against supplier-specific solutions for handling multiple screens, and use instead standard X protocol for communicating with connected screens. This has several important advantages: (i) Any type of screen can be connected, as long as it runs X. We have used Mira in a network with Unix computers, X-terminals, Macintoshes and PCs. The most cost-effective solution can be selected according to the need, for example a high-end 21" X-terminal with 1280x1024 resolution, or a low-end SVGA PC running X server software. Currently Mira only supports 8-bit screens, but this limitation will be lifted in the near future. (ii) The price of X-terminals is falling rapidly, and the more screens the lower the cost of each. A vendor-specific multi-screen solution on the other hand is likely to cost more per screen for many screens, due to the need for specialised hardware for extending the computer bus. (iii) An X-terminal can be located anywhere in a network, so that teleradiology can be made an integral feature of Mira. Since learning and user acceptance are important aspects of teleradiology, we found it imperative to have the same software serve both local needs and teleradiology.

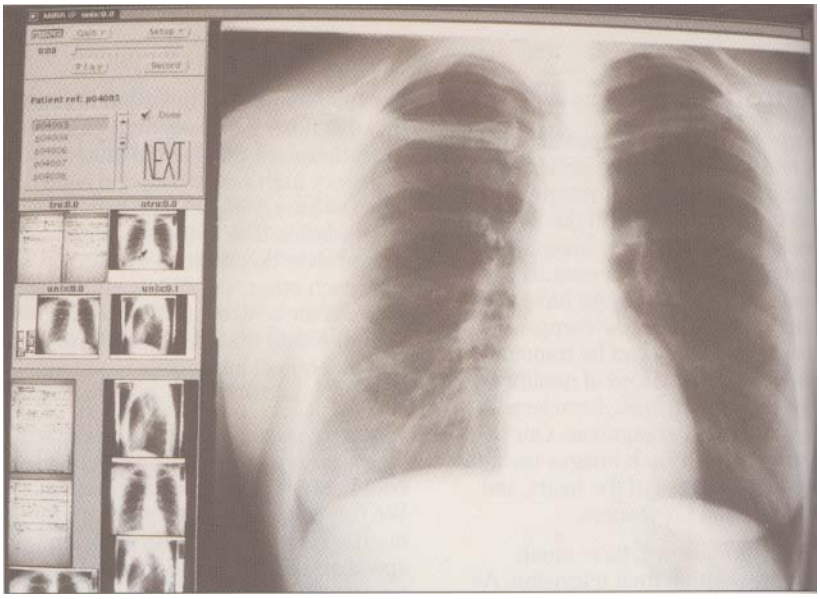


Figure 3 Close-up of console screen, with miniature images, four screen maps, image copies in each of the screen maps, and a full-sized chest image in the console

3 Working with many images

A patient examination has many images: A chest X-ray routinely contains both front and side projections. A complete spine investigation can contain 10 images from the skull and down. A CT scan will often have 20-30 transverse sections. Adding to this is the need to view previous images from the same patient. Even with a 4-screen system it may not be possible to display them all at the same time. How should the doctor select the one to display at full scale on the screen?

We rejected a text-based system for selecting images. The simplest system would give the doctor little information besides "image 1, image 2, ...". To be useful, each image would therefore have to be annotated with date, type of investigation, body part, projection, etc. The annotation would have to be done by the operator who digitises the images, requiring extra skill, and adding a time-consuming and error-prone step to the digitising process. For these reasons we chose instead to use miniature representations of the images, which all fit in the area of the console screen. The miniatures are generated automatically, and give the doctor an immediate overview of the images in an examination. The idea of miniaturised images is not new, in the literature they are called alternatively "thumb-nails", "stamps" and "icons".

The idea of miniaturised images fits in very well with Mira's concept of miniaturised screen maps. We think of the miniature as representing the original, whereas the full-size images are copies. A copy is made by dragging the miniature with the mouse to one of the screen maps, and the copy appears in the map as well as on the corresponding real screen. When the user clicks on a copy, the original is revealed by reversing its pixels. Vice versa, when the original is clicked on, any copy (or copies) in a map is highlighted. Figure 3 shows an investigation with 2 scanned requisitions and several chest images (in miniature), some of these are also displayed in full size.

Currently the operations implemented on a miniature are highlighting and copying to a screen map. Later, we plan to allow the same (reversible) image manipulations on the miniature as on the screen copy, this will be stored as parameters with the image and then reapplied every time the investigation is fetched from file.

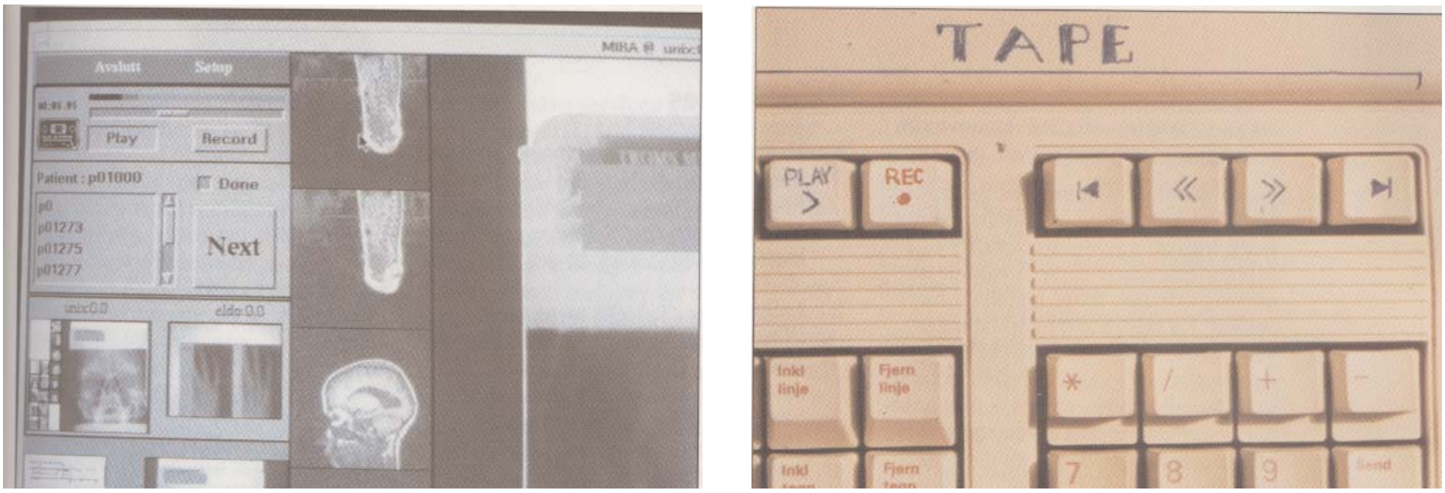


Figure 4 (Left) Close-up of console screen, showing interaction buttons, slide bar, and peak meter for the simulated tape recorde;. (right) showing tape functions accessible from the computer keyboard, without using the mouse

4 Integral sound

When the radiologist views images, he uses a dictaphone for taking the dictation. To discuss a patient case with a remote expert, the telephone serves for communication. In both cases the sound is treated separately from the images. In Mira we exploit the sound capabilities of modern hardware to integrate sound in a natural way with the investigation, alleviating the need for either dictaphone or telephone (3).

When Mira is used for local viewing and diagnosis, a simulated magnetic tape recorder is available. It is accessed alternatively from the keyboard or from mouse menus, see figure 4. Standard tape functions are simulated, like start, stop, record, fast forward and fast backward (with cueing). A time counter shows the current play position directly in seconds, and a horizontal position pointer shows the play position relative to the total recording. A horizontal bar indicator shows the instant play or recording level.

The simulated tape recorder has some special functions that set it apart from its mechanical counterpart. The "tape" can be instantly positioned to begin, to end, and to a previously set marker. The play pointer can be repositioned anywhere within the recording by dragging the pointer with the mouse. We have also experimented with stopping the recording during intervals with silence, this is a feature that is now appearing in high-end dictaphone units.

The dictation is stored as a digital sound file together with the images, hence tape loss or mixing of patient dictation is not possible. The dictation can be played back on the same computer, or later on a PC for transcription to text by a secretary. We are now contemplating automatic or operator-assisted voice recognition of the dictation as research projects, this may well be a reality in the near future.

When Mira is used for teleconferencing in a network, a simulated telephone connection is established by simply dragging a telephone symbol to the map of the networked screen. The transmission over the network uses standard Unix communication mechanisms, but since no high-level audio server standard exists like the X server standard for images, Mira depends on its own processes on each end to send and receive the sound.

Under optimal conditions the sound over the network is as good as or better than telephone. The connection is easy to establish, and with microphone and loudspeaker connected to the computer, the operator is relieved from holding a telephone while using the mouse and the keyboard. In addition to an ordinary one-to-one connection, other connections like one-to-many and many-to-one can established.

The simulated phone has some drawbacks, however. The connection uses the same physical channel as the images, so while an image is being transmitted the sound may be impaired. Also, since there is no echo cancelling, one must use a direction-sensitive microphone, pointing away from the loudspeaker.

5 Teleradiology installations based on Mira

Our first installation of Mira is now in daily use by radiologists at the University Hospital of Tromsø for viewing and diagnosing X-ray images taken at a small clinic 170 km away. The images are transferred over a 64 kbit/s line, this takes about 9 hours/day. The image transfer is implemented as an unattended process, and as soon as the images for one patient have been received, the corresponding patient number appears in Mira's list of incoming patients.

The communication in our first installation is based on traditional protocols and equipment: FTP transfer on TCP/IP, Ethernet routers, standard leased line. This has resulted in a very stable communications service. Over a period of currently 6 months of continuous use, more than 20 GBytes of image data have been digitised, transferred and viewed.

To test PC-based ISDN IP routers under heavy load and realistic usage conditions, we are now switching to ISDN for image transfer (keeping the leased line as a hot standby). Even when connected up to 10 hours/day, the total communications cost is expected to go down with ISDN as compared to a leased line. Image compression may further reduce transmission times from minutes to seconds, with a concomitant decrease in cost. Most important perhaps is the flexibility that ISDN will offer: Instead of sending the images to a fixed partner, the recipient can be selected dynamically according to expertise, load, etc.

In our second trial based on Mira, hospitals are connected over a dialled network with an aggregate capacity of 2 Mbit/s. Currently three hospitals are on-line, with two more installations coming shortly. Each hospital has a Mira workstation, and the digitising equipment comprises a Peltier cooled, CCD-based image camera, a film scanner, and Computed Tomography (CT) image converters based on diskette transfer.

The digital communication is implemented on a DIAX "Switched Wideband System" (SWS) network. The dial-up connection is established on demand by software, whenever an IP address that has a corresponding SWS number is requested (4). The point-to-point connection is almost transparent to the user and to the high-level programmer, who see only a network number. Almost, because it is not currently possible to establish more than one logical connection at a time on one physical line. This will be remedied when the latest release of the DIAX SWS is installed. It will then be possible to split dynamically the 2 Mbit/s channel into a number of separate n x 64 kbit/s connections.

6 The future of teleradiology

Mira has proved itself as an efficient tool for the use of telecommunications in the medical field. We have demonstrated the possibility of using teleradiology on a routine basis for medical diagnosis, thus extending a timely radiology service to a remote location. We also believe that our field trials with teleradiology over a high-speed dialled network mark the beginning of a truly distributed radiology service, where one can search expert advice, exchange ideas, and communicate irrespective of geographical location.

References

1 Mun, S K et al. 1992. Command-wide Teleradiology for US Armed Forces in Korea. *Proc. Medical Imaging VI: PACS Design and Evaluation.* SPIE 1654, 81-88.

2 LaStrange, T. 1990. swm: an X Window Manager Shell. In: *Proc. Summer USENIX Conf.*

3 Sund, T, Kileng, F, Størmer, J. 1992. Integrating voice with images and text in a workstation for radiology. *Proc Medical Imaging VI: PACS Design and Evaluation.* SPIE 1654, 563-565.

4 Jøkulsson, G S, Kongsli, T A. *Providing Transparent Internet Connectivity Over GLOBAND Circuits.* Tromsø, NORUT Information Technology. (Norut-IT Report 3017-2.)

Telediagnosis in the context of digital image analysis

BY HÅVARD E DANIELSEN

621.39:61

The use of telecommunication technology to assist in the delivery of health care, is a solution of increasing popularity, especially to some of the problems faced by rural residents in obtaining health care. Many consider telemedicine or medical telematics as a solution to the problems of delivering health care to remote areas and areas underserved by clinicians.

The noble approach

The idea behind telemedicine is said to be to ensure a health service to all people regardless of their situation. Access to medical services may be limited by geography, climate, communication, transportation and economy, as well as shortage of trained personnel. Telemedicine has proven to be effective as a disaster response (e.g. the earthquake in Soviet Armenia in 1988), a meaningful aid for the third world (e.g. the telemedicine link between Canada, Kenya and Uganda), care for the elderly (e.g. security alarms) and chronically ill in their homes (e.g. home based pulmonary function monitor for cystic fibrosis), as well as a provider of health care to remote or isolated areas.

The practical approach

Telemedicine is also to a large extent a practical and economical solution to a number of problems associated with health care. There is a lack of medical specialists in many fields, such as pathology and radiology, and it is further more not economical to provide a number of advanced medical services at the smaller hospitals. Instead of having the specialists or the patients travelling around, a long-distance telediagnosis will save money and time. Such activities will also help to reduce the isolation of rural health care professionals, as well as provide an opportunity for continuous medical education and training. The possible benefits from telemedicine are numerous, as are the fields of application.

Telepathology

A good example is found in telepathology. Pathology is the science of diseases in tissues and organs, and the pathologist provides a diagnosis based on microscopical examinations of cell and tissue sections. Hence, telepathology is the same service, but carried out remotely by means of transferred microscopic images of tissue samples displayed on a screen. Only the largest hospitals have services in pathology. Smaller hospitals and clinics normally deliver surgical biopsies and cell samples to these hospitals or commercial laboratories for examination and diagnosis. Thus, in a number of cases the patient has to be operated on twice, first to obtain a biopsy for morphological examination, and later for curative surgery. One of the special diagnostic functions of a department of pathology is the frozen section service. This is a rapid diagnostic procedure carried out on fresh tissue. The tissue is frozen, cut, stained and examined within 20 minutes, and the importance lies in the fact that the tissue analysis can be performed while a surgical operation is in progress. The continued surgical procedure is in certain cases dependent on this type of pathology service, and hospitals without access to frozen section service cannot treat such patients, but has to refer them to another hospital. These are typical cases were use of telediagnostics would strongly benefit the patients, and at the same time reduce costs and free medical resources.

An old concept

The concept of telemedicine is not novel. In historical terms the foundations for medical telematics were laid in Italy in 1935 with the creation of the International Radiomedical Centre in Rome. A number of pioneering telemedicine programs were initiated in USA and Canada in the 1960s and 1970s, and during the last two decades more than a hundred medical papers have been published on the use of medical telematics in health care and diagnosis. Medical images and other data are now transferred on-line in real time between local health personal and consultant physicians and other specialists in central health institutions all over the world.

Telemedicine comes of age in Norway

Although the concept is old, the wide acceptance and use of telemedicine is yet to come. Most patients have never been confronted with this technology, and indeed very few physicians have ever taken part in a telemedical consultation. This is also the case in Norway, although the status here seems to change rapidly. There is a growing interest for telemedicine in Norway, primarily due to the pioneering telemedicine project of the Norwegian Telecom and the University Hospital of Tromsø. A number of telemedical services have been developed and tested between the University Hospital of Tromsø and two local hospitals in northern Norway, and the University hospital are now offering telediagnosis on a permanent basis for specialised services in pathology, radiology and dermatology.

The modest approach

Current telemedical technology benefits from recent developments such as the decreased cost and improved quality of the coded-decoder (codec) equipment use in interactive digital video systems and the expansion of fibre-optic cable networks.

In contrast to other telediagnostic systems, the Norwegian group has constrained their design of telemedicine in Norway to the specifications of standard available telecommunication systems. The telepathology system was designed to use the 2 Mbit/s capacity of the "mega-net", a band width corresponding to approximately 32 telephone channels (i.e. 0.1 MHz). This represents one hundredth of the capacity of the microwave channel employed with the Corabi DX-1000 Telepathology system at present use in Atlanta. Dr. I. Nordrum and co-workers at the telepathology project in northern Norway have later shown in a limited study that images with sufficient quality for frozen section service may be transferred at 384 kbit/s capacity. This indicates that telepathology in the very near future may be carried out on the standard ISDN net (with three standard telephone lines, each two x 64 kbit/s, connected with an inverse multiplexer). The reduction in costs is drastic, and probably what is needed for a widespread use of such systems.

The real challenge

Pathologists at the Emory University Hospital and the Grady Memorial Hospital consult one another on a fully bi-directional point-to-point microwave telecommunications system that link the two hospitals. They are transmitting high resolution full colour images at the rate of 30 images per second, a transfer speed that enables full size full quality image viewing in real time. They have, as in Tromsø, the ability to manipulate the microscope at the other hospital, and view the microscopic slides on monitors in their own laboratory. In contrast to the Tromsø group, they have no image degradation when the microscopic slide is moving on the microscope stage.

A high resolution full colour image from a microscope (e.g. with a size of 1024 x 1024 pixels and 8 bit resolution per pixel per colour) represent roughly 3.1 Mbyte of data (-25 Mbit). The capacity needed for transmitting 30 such images per second in an uncompressed form is thus 750 Mbit, and the cost is significant (and too high for the benefits obtained, at least in Norway). On a 384 kbit/s line each image (in uncompressed form) would need a little more than one minute for transfer, and hence, living images in high resolution and full colour is out of the question.

The technology needed for telediagnosis in pathology (and other disciplines as well) is already available. The real challenge is to make such systems work as close to the standard methods as possible, employing telecommunication systems available on a general basis, and with acceptable quality and costs. The Norwegian Telecom and the University Hospital of Tromsø seem to have taken on this challenge with their project "Telemedicine in North Norway", and definitely gone a part of the distance. However, as they say, the best is probably yet to come.

Different strategies

There are two principle strategies to follow, either to increase the transfer capacity without increasing the costs, or to increase the transfer speed by reducing the amount of data to be transferred. As long as 750 Mbit lines are not likely to be a part of the telecommunication system generally accessible to the public in the near future, the latter strategy appears to be the more valid one. A reduction in data may be accomplished in several ways:

- reduction in image resolution
- reduction in image size
- compression (packing) of data
- processor controlled transfer schemes (e.g. to transfer only those pixels which are different from one image to the other).

Although all these methods will reduce transfer time, a reduction in image quality is also inevitable. The more effective reduction in time, the more significant image degradation. For some applications, a significant image degradation is acceptable, for others it is not.

Different applications – different images

Different applications do not only have different acceptance levels for image degradation, but they vary in initial resolution, image size, colour versus b/w and static versus living images. Several different features might be important in an object, and of special concern to a given application:

- mobility
- architecture
- morphology
- colours
- density
- intensity
- texture.

Mobility clearly demands living images, and real time transfer is also important in e.g. endoscopy (examination of body cavities with a camera). Real time transfers are also important in telepathology during the selection of the putative diagnostic fields in tissue specimens, but having located those fields, transfer of static images are sufficient for detailed examination. High resolution and full colour are important for such examinations of standard prepared specimens, but resolution and colours may be reduced for freeze section service as these specimens are low in details and growth pattern is the prevailing feature analysed. Radiographic images have clearly no need of real time transfer, but depend heavily on large size images of high resolution for analysis of grey level distributions and density. Typically a radiographic image comprises 4096 x 3584 pixels with 12 bits resolution/pixel. This represents roughly 22 Mbyte of data (1.76×10^8 bits), and the transfer time on a 64 kbit/s telephone line would be nearly 46 minutes. So, even static images need high capacity telenet and/or advanced compression algorithms.

Since functionality as well as the need for quality differs from one application to another, it might be profitable to treat these different images individually. For some applications the morphology and architecture are the most important features (e.g. freeze section service), for other (e.g. X-rays) grey level distribution and density prevails. The need for individual treatment of images makes strong demands on the methodology, but the fact that different features are important in different applications also opens new possibilities with regards to data reduction.

Alternatively (or actually supplementary) to the different algorithms for data compression one could look for methods to enhance image validity before and after transfer. Image validity in this context means those regions in the image that is of special interest. To put is simply, if the objective was to evaluate the size of the nuclei within the cells, the nuclei could have been segmented and the microscopy image reduced to a binary (two bit) image before transfer. A more realistic example would be to enhance contrast or reduce noise in order to maintain image validity during e.g. compression. The methods of noise reduction, image enhancement and segmentation belong to the discipline called image analysis.

Digital image analysis

Image analysis is a set of standard techniques with wide applications in all branches of science and technology. Beyond a visual examination of an image, there are several methods of further analysing the image. The objective is to extract the maximum amount of structural information, and image analysis provides a quantitative way of assessing image defects and image resolution. Based on such analyses, structural and non-structural (noise) information may be separated to give an enhanced image, which shows more clearly the structure of the objects which are of interest. Clearly, if that is achieved, one can afford more image degradation due to data reduction, and still maintain the necessary qualities for the final visual examination (or quantitative analysis). Correspondingly, image processing techniques may be applied on the image after decoding to compensate the loss of resolution due to e.g. compression. Further more, quantitative image analysis may ease the burden of long distance diagnosis by extracting new objective features from the image.

The concept of image analysis is data reduction, and the principal steps are outlined in figure 1. This data reduction is however directional, in the sense that one ideally keeps the data of interest and discharge the rest.

Image capture

The first step is to convert the image into an electronic signal suitable for digital processing and storage. An image is a 2-dimensional distribution of energy, typically of visible electromagnetic radiation (light) but can also be of x-rays, ultraviolet, infrared or other radiation; electrons, acoustic waves or even nuclear particles. For most purposes in telediagnostic one deals with visible light images captured by television-type cameras. The task of the image capturing process is to quantify the image in both space and tone. Spatially, the image is divided into a square or rectangular grid of picture elements, known as pixels. Tonally, the image is separated into intensity levels, e.g. 256 levels in an 8-bit (2^8) resolution.

Image enhancement (tone processing functions)

The next step would be tone (e.g. grey tone) image processing for image enhancement. These functions are typically used for contrast enhancement, noise reduction, and other filtering processes. They are grouped into linear and non-linear functions, as well as frame & window operations, pixel operations and matrix operations (see figure 2). With linear functions, the same algorithm is executed for all pixels, whereas non-linear functions are those were different operations and algorithms can be performed for the pixels depending on the local tone information. Frame & Window operations are global functions working on whole images or windows, and the tone values are not changed by these operations (e.g. zoom, shift, rotate, scroll). In pixel operations, all pixels are processed individually by mathematical calculations without influence of the neighbouring pixel information. The opposites are matrix operations, were every pixel is calculated regarding to the grey value information of its neighbouring pixel up to a defined distance or matrix.

Segmentation

Segmentation is the function that separates the regions of interest (ROI) within an image from the background. ROI or objects can be segmented by:

- grey value
- colour
- edges
- texture.

There are three main methods of segmentation; i) thresholding, ii) edge finding, and iii) region growing. Segmentation by threshold / discrimination is the simplest and most common method. At its simplest, all tones below a selected level are treated as of interest and all above as background. One may use more than one threshold, allowing several scales of grey tones or colours to be of interest. Thresholding works well in situations where the illumination can be carefully controlled to have the same level across the entire scene, as e.g. in microscopy. A more adaptable technique is edge finding, which detects the regions of high rate of change of tone in an image, on the basis that these are likely to indicate edges of objects or ROI. The third approach looks at neighbouring pixels and group them together if they are sufficiently similar.

The segmentation process generates a binary image (each pixel has only one of two states; ROI or background) which is used as a mask for further processing of the original image.

Measurements

Having a digital and enhanced image with the regions of interest identified, objective quantitative analysis or measurements can be performed. Such measurements may be used to strengthen or improve the subjective diagnosis, and a number of new features may be added, such as volumetric estimations in ultrasound images, sub-visual objects in x-ray images and DNA content in tumour cells to mention a few.

Pre-transfer methods

As indicated above, there are several ways to improve the quality (or validity) of an image, and the choice of methods will depend heavily on the type of image and application. If we stay with the examples of telepathology and light microscopy, there are four methods which are especially helpful:

- contrast enhancements
- shading correction
- noise reduction
- haze removal.

Contrast enhancement functions are based on pixel operations and the main applications are contrast manipulation to reach better visibility of structures in certain tone ranges.

Shading correction is a function that is used when inhomogeneities in the background (or sometimes also in objects)

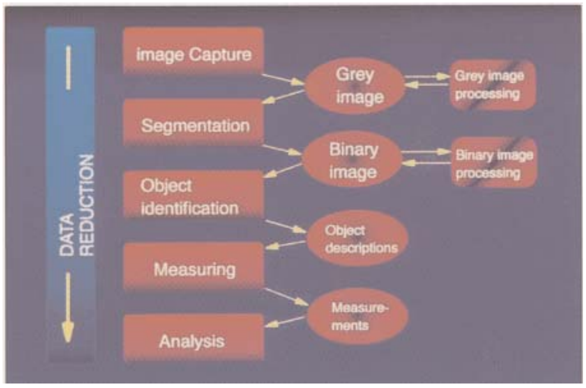


Figure 1 The principles of image analysis

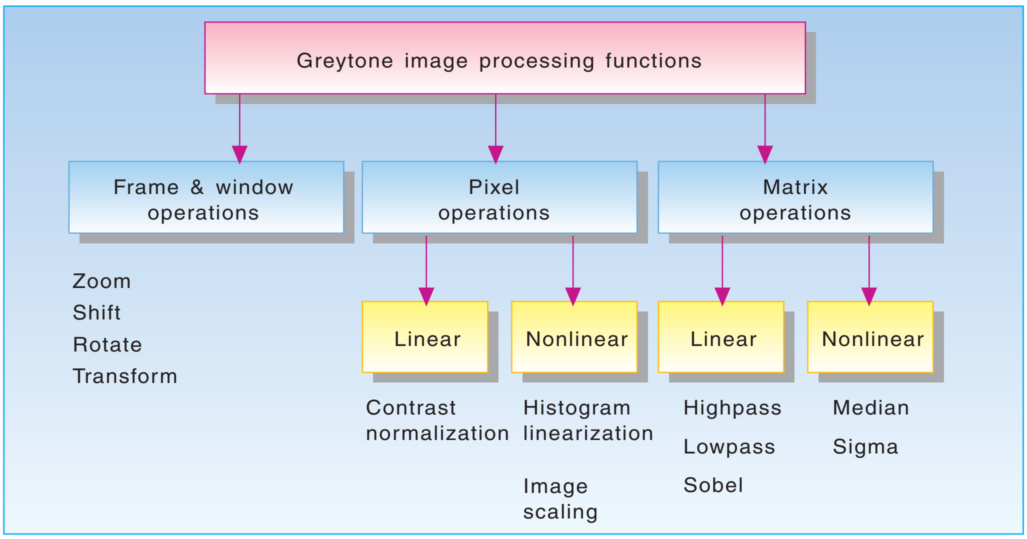


Figure 2 Different methods for image filtering and image enhancement

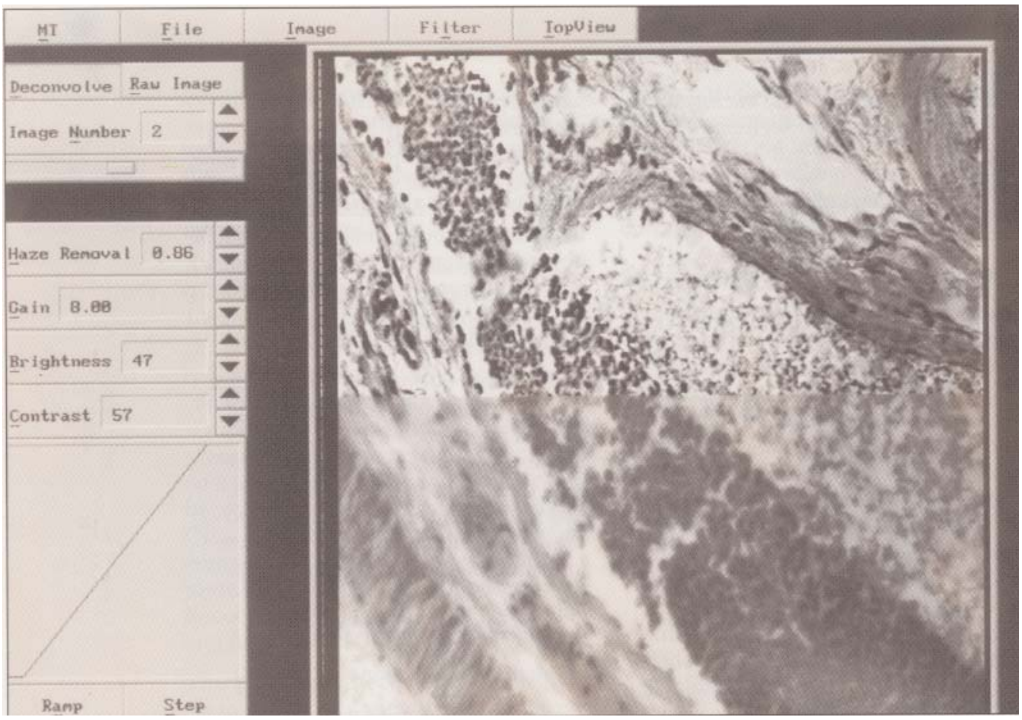


Figure 3 An example of deconvolution with the nearest-neighbour algorithm performed on a microscope image of a histological section using the Micro-Tome system

Figure 4 Enhancement of fluorescence signals by deconvolution technique

reduce the "sharpness" of an image. In light microscopy this may be caused by uneven distributed light through the object or by dirt on the camera or lenses. The simplest way to correct such shading effects is to use an empty field image (which contains only the shading effects) as a reference image and thereby remove the contents of this image from the real image. If no empty field image is available, an artificial one may be generated from the original one by filtering it with a low pass filter (with a large matrix size, e.g. 100 pixels).

Most images are blurred by a varying amount of noise from light sources, camera etc. Noise can be removed by analogue as well as digital manipulations of the image. The most common filter for digital noise reduction is the median filter or the sigma filter. Because noise is a statistical event, the sigma filter calculates the standard deviation in a matrix and smoothes pixels which deviate abnormally from the mean tone value.

The most drastic improvement of an image from a microscope is obtained by confocal techniques. The image seen through a microscope includes the in-focus portion and the out-of-focus portion above and below the plane of focus. The haze or blur produced by the out-of-focus planes is a natural consequence of the optics of the microscope, but can be removed by so called deconvolution algorithms. These algorithms have the same function as the apertures in the laser scanning confocal microscopes, – removing the out-of-focus portion of the image. The most common is the nearest-neighbour algorithm, a very fast procedure that may be applied either before or after the images have been transferred.

Post-transfer methods

Contrast enhancements and noise reduction may also be applied to the images after transmitting them through the telenet. The transfer process will generate new noise that may be removed, and reduced contrast due to the compression/decompression algorithms may to a certain extent be compensated for by digital post-processing. Manipulation of grey values has proven to be most valuable during evaluation of radiographic images where the resolution has been reduced in order to reduce transfer time.

Old methods – New services

The different projects concerned with telediagnosis have so far, for obvious reasons, been constrained to the traditional ways of diagnosis, namely subjective observation of the specimen or image hereof. During the last decades a number of new diagnostic and/or prognostic methods have been developed. This is especially true for pathology, where a number of techniques combining traditional pathology and fields like molecular biology, immunology and/or image analysis have evolved. For most of these techniques, the preparation of specimens for analysis is simple and can be carried out in any laboratory without specialised equipment or expertise. It is usually the evaluation or analysis that require particular skills or advanced equipment, and hence, these methods would be suited for telediagnosis.

Tumour ploidy

Determination of the amount of DNA in solid tumours is of special interest because it gives prognostic information about patients with cancer in bladder, ovary, breast, kidney, lung and colon. Image cytometric (ICM) DNA analysis form part of the routine examination of patients with gynaecological cancers at The Norwegian Radium hospital and yield important and decisive information about treatment and prognosis. The tumour tissue is dissolved into single cells by means of proteolytic treatment and the resulting monolayer is hydrolysed and stained with the DNA-specific Schiff reaction. The ploidy is determined by means of quantitative image analysis, where integrated optical density becomes a measure for the DNA amount. This preparation is easily carried out by the local laboratory, but the measurement and analysis require trained personal and image analysis equipment. A telediagnostic approach would require approximately 20 random selected images with standard resolution in black and white (512 x 512 pixels with 8 bit resolution per pixel) which can be transferred with standard ISDN telecommunication within less than six minutes. The measurement and analysis would be performed off-line in the department of pathology within 30 - 90 minutes and the resulting DNA histogram and diagnosis may be transferred back as one image to the local laboratory.

Hormone receptors and antigen markers

A similar case exists for quantification of different hormone receptors as well as identification of other antigens by immunostaining applying monoclonal or polyclonal antibodies. Such immunohistochemistry methods have gained applicability in several special diagnostic cases, but the use is normally limited to larger central institutions since special experience and/or special equipment is required to analyse these preparations. Again, local hospitals could gain access to this experience through the telenet.

Karyotyping

Chromosome analysis is another example of an important diagnostic method where the utilisation is limited due to lack of specially trained personal. Although the preparation of suitable metaphases is a little more demanding compared to the methods mentioned above, it would be advantageous to do this locally since the preparation requires living cells. The main obstacle for a widespread use of karyotyping as a diagnostic method is the fact that it takes 4 - 6 years of training to be able to set up the correct karyotype. Most cytogeneticists employ digital images and computers for the arrangement and analysis of chromosomal preparations anyway, and the method should therefore be well suited for telediagnostics. Again the requirements for image size and resolution are low, and a 768 x 640 b/w image with 8 bit per pixel is more than sufficient. Ten metaphases could therefore be transferred with a standard telephone-line on the ISDN-net in five minutes. Transmittance in real time or on-line analysis is not required unless a consultative dialogue is desirable.

Quantitative image analysis in medicine

The use of morphometric analysis has a long history in pathology, were measurement of nuclear size for histological grading and counting of mitotic cells as a prognostic indicator are typical examples. Image analysis is more recently also used for automatic classification of tissue specimens, and there exist several systems for automatic pre-screening of cytosmears (cell samples).

Besides the examples from pathology, quantitative image analysis is used in a number of diagnostic methods in medicine. Image analysis is a prerequisite for such diagnostic methods as computer tomography (CT) and magnetic resonance (MR), and is frequently used as part of diagnostic apparatus in ophthalmology and cardiology. 3-dimensional reconstruction and visualisation are other methods which are fast growing in medicine, and where image analysis is the foundation.

The image analysis systems are often expensive add-on equipment to standard apparatuses, which in addition to costly investments requires specially trained personal. The link is either a digital communication or an analogue camera, and it will therefore in many cases be possible to examine the patient with the standard apparatus at the local hospital while the specialist is analysing the images at a central institution (e.g. volumetric heart analysis with ultrasound, classification of EEG results, enhancement of radiographic images, quantitative analysis of microscopical images etc.).

How to have the cake and eat it too

One of the main controversial issues regarding national health care has been whether to centralise the hospital treatment of certain diseases such as cancer. Arguments in favour of centralisation have been to save money and obtain full use of expensive equipment, as well as improved quality of treatment due to greater experience. This is compared against the advantage of being treated in the local hospital.

By telemedicine and image analysis, the possibility exists that a large number of patients may be treated in their local hospitals by medical specialists in central institutions, and this might prove to be the solution of the national health administrators dilemma; how to have the cake and eat it too.

The scientific approach

Apart from the many practical and economical advantages gained by telemedicine, one should not overlook the scientific value of communicating digital images through the telenet. The digitalisation of images is easily combined with a database in which every case may be classified and commented on for future reference. Besides becoming an excellent tool for training, such databases may be used for feature extraction by image analysis systems. The extraction of features is the basis for an objective description of the cases, and if that is obtained, further possibilities are automatic recognition and classification systems.

Communicating digital images in a consultative function will by definition promote medical education and training. The very existence of telediagnostic equipment will stimulate co-operation between clinicians and researchers in different institutions, as well as nourish development of new applications.

Whether telemedicine will be a general success will to a large extent depend on whether or not one manages to meet the real challenge – to make such systems work as close to the standard methods as possible, employing telecommunication systems available on a general basis, and with acceptable quality and costs. Current telemedical technology benefits from recent developments such as the decreased cost and improved quality of the coder-decoder equipment, and such equipment will be further developed in the near future. By an additional implementation of image analysis, telediagnosis will continue to grow, both in distribution and application.

Teaching and learning aspects of remote medical consultations

BY SIGMUND AKSELSEN AND SVEIN-IVAR LILLEHAUG

621.39:61

Abstract

This paper discusses educational effects of remote medical consultations. First, a definition of telemedicine and a brief overview of some of Norwegian Telecom Research's remote consultation applications are given from an educational perspective. Some state of the art theories of teaching and learning are presented and discussed. Then, the focus is put on functionalities that can increase the educational outcome from remote consultations. Finally, possible impacts on society from life-long learning through remote consultations are discussed and thoughts on future work are given.

This paper originally appeared in: Teleteaching. Proceedings of the IFIP TC 3 International Conference on Teleteaching '93, Trondheim, Norway, 20-25 August, 1993 (G Davies, B Samways, Eds.). IFIP Transactions. Elsevier Science Publishers BV (North-Holland).

1 Introduction

In 1988, Norwegian Telecom Research (NTR), in co-operation with the University Hospital of Tromsø (UHT), started a project on telemedicine in North Norway. North Norway is characterised by sparsely populated communities spread over vast distances. Furthermore, the increase in complexity and amount of medical knowledge has caused a demand for specialisation within the health services. As a consequence there is a limited number of qualified personnel within certain sectors of the health services, especially in remote areas. An objective of the project is to use telemedicine to provide equal health care services to each individual in Norway, regardless of geography or economic variation in the population.

Telemedicine is not a new concept. Health care professionals have been using the telephone to carry out their services for years. In addition, research efforts have been put into utilising more of the telecommunication repertoire including speech, text, data, picture and video communication (8). A goal for telemedicine is to eliminate travelling for patients and specialist. Telemedicine may be defined as (2):

> *The investigation, monitoring and management of patients and the education of patients and staff using systems which allow ready access to expert advice and patient information, no matter where the patient or relevant information is located.*

The NTR telemedicine project systematically explores the possibilities provided by new information and communication technologies in improving health care services, and has led to several telemedicine applications. Today, these applications are based on a variety of networks, ranging from the ordinary telephone network to specialised data and videoconferencing networks. The applications may be grouped into four categories (3): remote diagnosis, distance learning, medical information and administrative information.

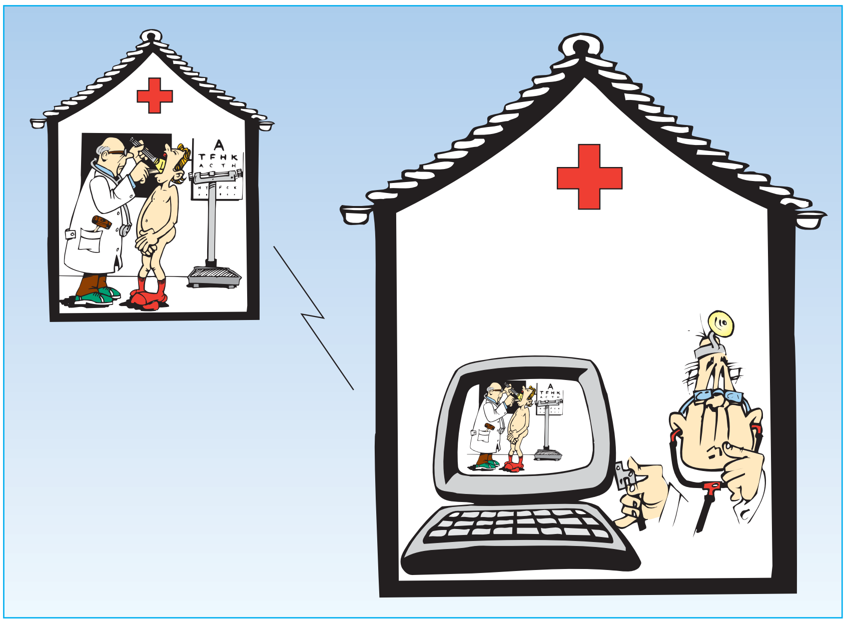


Figure 1 General set-up for remote consultations (3)

In this paper we focus on remote diagnosis as collaborative work and discuss the educational effects it has on the health care personnel in remote areas.

2 Remote diagnosis

The term *remote diagnosis* refers to remote consultations where the diagnosis is based on medical data transmitted over telecommunication networks. Usually there are three persons involved in a remote consultation (see figure 1). The most important is the patient concerned about his disease. There is also the general practitioner who has insufficient knowledge to assess the patient's problems, and finally there is the specialist who is supposed to diagnose the problems and propose treatment if possible.

Most of NTR's telediagnosis applications are based on transmitted medical pictures, both live and still. The applications include, among others, remote consultations within: pathology (15), radiology (21), endoscopy (7), dermatology (10,12), cardiology (1) and even psychiatry (9). In addition, trials have been conducted on remote consultations within microbiology, gastroscopy, electrocardiography and electroencephalography. All these applications are similar in many aspects, both in their technological solutions and methods of work, but different in others.

A majority of the applications utilise videoconference equipment over a 2 Mbit/s switched network. Some of the applications include add-on special purpose equipment such as a microscope or endoscope to provide "the necessary pictures". Most applications demand all involved persons to participate at the same time (synchronous communication). Lately, research and development has begun in order to provide several of the applications on lower transmission capacity and, thereby, cheaper networks. ISDN is, in this respect, seen as an obvious network solution (3). Terminal equipment used in the ISDN solutions include the Tandberg Vision videophone (14), and NTR's own system for digitising, storing and transmitting medical stills (pictures) from arbitrary video sources. Further, the applications will also provide for asynchronous communication, (i.e. the participants will not necessarily be present at the same time).

Below, tele-endoscopy, remote diagnosis of skin diseases and remote echocardiology will be described in more detail with focus on the educational aspects. Below, the term practitioner will refer to both general practitioners in primary health care and residents in remote hospitals.

2.1 Tele-endoscopy

An endoscope is a device for guiding a source of light inside the human body and transposing an image of the examined organ on a monitor. Tele-endoscopy refers to the transmission of these images over the telecommunications network to a specialist at the other end. A practitioner is, however, still needed to bring the endoscope inside the patient and adjust the positioning to get the correct images (7). Tele-endoscopy is used in North Norway for ear, nose and throat (othorhinolaryngology) consultations. Telecommunications are presently based on a 2 Mbit/s connection. Videoconferencing equipment provides a two-way audio and video connection between the specialist and the practitioner in charge of the patient. As a service, tele-endoscopy is meant to cover two aspects: remote consultations and transfer of competence to the practitioner.

An endoscopy session is divided in three parts: examination, diagnosis, and compilation of a treatment plan. From an educational perspective, tele-endoscopy goes through four phases. First, the practitioner has to go through a short tutoring period together with a specialist at UHT. During this period he acquires the basic knowledge and skills that are necessary to operate the endoscope. In the second phase the practitioner is back at the remote hospital where he operates the endoscope with directions from the specialist. The specialist explains the endoscope movements he has requested and what is being seen. Through this process the practitioner receives practical instruction in endoscopy examination techniques. Based on the findings from the endoscopy and through conversation with the patient, the specialist makes his diagnosis and plan for treatment, all while explaining his reasoning to the practitioner. In the third phase the practitioner begins to more actively cooperate with the specialist in the examination, diagnosis and compilation of a treatment plan. Finally, in the fourth phase, the practitioner is ready to take over the consultation. Remote consultations are only used for patients that the practitioner finds difficult to diagnose. At this stage the specialist monitors and evaluates the practitioner's examination, diagnosis and proposed treatment plan on a limited scale.

2.2 Remote diagnosis of skin diseases

Remote diagnosis of skin diseases is based on technical solutions similar to tele-endoscopy. The only difference is that instead of the endoscope the practitioner has several cameras which are used to transfer still and motion pictures of the patient to the specialist. As with tele-endoscopy, a session of remote diagnosis of skin diseases can be divided into examination, diagnosis and compilation of treatment plan. From the educational perspective the service can be described through three phases. First, the practitioner functions as the specialist's extended arm by providing the specialist with different pictures (by moving and focusing cameras) of the patient, and giving comments about how the skin feels as well as other necessary anamnestic information. This process might also include the practitioner taking a biopsy or performing other tests. Throughout the session the specialist explains the different actions as well as the reasoning behind his diagnosis and the resulting treatment plan. In the second phase the practitioner and the specialist cooperate as colleagues during the consultation. As the practitioner gains the necessary knowledge and skills through collaborative practising, he is ready to move over to the third phase of independent consultation with the patient. Similar to the tele-endoscopy service, the specialist is brought in whenever necessary to help, monitor and evaluate and the practitioners' actions.

2.3 Cardiology diagnosis through remote echocardiology

The project of remote diagnosis through echocardiology has taken a somewhat different approach than tele-endoscopy and remote diagnosis of skin diseases. So far the main topic has been to investigate the appropriateness of remote echocardiology as a tool for distance education of physicians with minor practical experience in echocardiology (1). The technical solutions are similar to the other presented projects. In addition to videoconference two dimensional, M-mode and doppler echocardiology are transferred.

The practitioner involved in the test trial had no further experience with echocardiology than a five days theoretical introduction course. During the test trial, 38 patients were first examined by the practitioner with the specialist remotely monitoring the practitioner's performance and giving supervision. Practical instruction in examination techniques, diagnosis and compilation of treatment plans were given by the specialist whenever required. To double-check the practitioner's performance, and to evaluate the quality of remote echocardiology, the patients were directly examined by the specialist later. Based on a comparison between the two consultations of each patient, Afseth et. al (1) concludes that remote echocardiology is an efficient and appropriate technique for education in echocardiology.

3 Educational implications of remote consultations

The projects described above have similarities both in technical solutions and phases of execution. A typical session goes through the steps of examination, diagnosis and compilation of a treatment plan. Within each discipline the practitioner evolves from having minor experience in the field to becoming a semi-specialist through three phases: practical demonstrations given by specialist, practising in co-operation with specialist and, practising under supervision of the specialist. This approach of learning is known as the *role model* within the medical system (13). The role model has its parallel in educational theory in what Collins et. al and Brown et. al (4,5,6) have proposed as the *cognitive apprenticeship* model for instruction of cognitive skills. Since ancient times apprenticeship has been the most common method to transmit the knowledge from a master to an apprentice in a wide variety of fields, also within medicine. In their argument for cognitive apprenticeship, Brown et. al claim:

> *"... that knowledge is situated, being in part a product of the activity, context and culture in which it is developed and used ((4), p 32)."*

The cognitive apprenticeship model comprises three methods similar to the three phases of the role model: situated modelling, coaching and fading. These are described as follows in (4):

- *Situated modelling:* The teacher/coach promotes learning, first by making explicit their tacit knowledge or by modelling their strategies for students in authentic activity.
- *Coaching:* The teacher/coach supports the student's attempts at doing the task.
- *Fading:* The teacher/coach empowers the student to continue independently while the support is gradually withdrawn.

Collins et. al (6) summarise the benefits of the cognitive apprenticeship model for instruction as follows:

- Learners will more easily understand the purpose and use of knowledge.
- Learners learn by actively using, rather than passively receiving, knowledge.
- Learners learn different conditions under which they can apply their knowledge.
- Learners can generalise the knowledge across meaningful situations (instead of directly acquiring a generalisation without any context).

A good educational outcome of the CA model relies on good and efficient communication between the apprentice and the specialist throughout the different phases of the model. Both the practitioners and the specialists involved in the three projects presented above, report a better educational outcome through remote consultations as compared to the traditional specialist education within the respective disciplines. They explain that the practitioner in a remote consultation receives much more attention and supervision from the specialist than a physician going through his special education at a central hospital. This observation corresponds with the following shortcomings identified in an evaluation of the traditional education in ventilator therapy among nurses and physicians specialising in intensive care and anaesthesiology at UHT (13):

- There are weaknesses in the supervisor's ability to explain cases/problems and their solutions.
- Feedback on student performance is missing or of varying quality.
- Students are rarely required to explain cases/problems encountered or their own solutions.
- Supervisors are not available when needed, and when available, time limitations prevent the supervisor from demonstrating or giving detailed enough explanations.

A remote consultation provides an environment that, by its characteristics, forces the participants to bypass these possible weaknesses in traditional education. These characteristics can be identified as:

- The specialist needs to involve the practitioner in the procedures in order to carry out the examination/consultation.
- The specialist is dependent on the practitioner doing a good job.
- The practitioner is more frequently required to explain his reasoning and performance.
- The specialist is available during the entire session.
- The specialist saves time as the practitioner begins to work more independently.
- The specialist can use his time on challenging cases, as more patients are treated by the practitioner.

As a result, both the practitioner and the specialist are highly motivated to follow the CA methods of situated modelling, coaching and fading thereby leading to an efficient educational outcome of each session.

A measure of the success of the NTR's telemedicine project is the high number of requests from remote hospitals and health institutions that want access to the different services. The hospitals want to take advantage of the benefits that telemedicine provides. For example, the patients can receive specialist consultations locally. In a remote consultation, the three participants have the opportunity to discuss the problem at the same time. This has not been the case earlier. The gap in time from the problem's occurrence, the call of the practitioner's attention to it, the specialist consultation, and the practitioner being informed of the diagnosis and the treatment plan, is reduced. Same time, same subject and same information available is, according to Jøsendal et. al (12), the remote consultations' greatest pro. The medical personnel (in primary health care) get better opportunities to keep professional contact with other institutions and groups of specialists. This may improve stability, competence and recruiting of medical personnel, thus increasing the possibilities for running a qualitatively good and safe service. Further, through good telecommunications the medical personnel can take part in an exchange of knowledge and professional discussions independent of geographic location. The personnel can raise their competence and get further education without leaving their district and thereby save many and expensive travels. Time saved on travelling can instead be invested as work time. In summary the remote areas can provide a better health care system at a lower cost.

Up until recently, most of the services in the telemedicine project have been performed on a trial basis to a few remote sites. Opening the services for more widespread use will increase the load on the specialists and their departments at the central hospitals. Therefore, the different services need to be investigated for how to most efficiently meet the new demands.

4 Increasing the educational outcome

A main focus of the telemedicine project has been to provide remote medical consultations by combining advanced medical technology and telecommunications. The above discussion implies that the educational outcome of remote consultations might be just as important as the service provision. This has lead to applications, such as remote echocardiology, where the main topic has been to investigate the appropriateness of remote consultations as a method for distance education of health care workers.

One way to prevent an increased load on the specialists and their departments might be to make the practitioners more independent. This can be done by providing additional tools for educational purposes and specialist advice. In other words tools that offer the practitioner an efficient learning environment combined with supervision functionalities without necessarily having access to the specialist. Computer applications have the potential to facilitate both learning and assistance during a consultation. By inte-

Table 1 Shuell's learning functions with examples (20)

Function	Teacher initiated	Learner initiated
Expectation	Provide overview (map, diagram); Statement of purpose	Identify purpose for using the program
Motivation	Opportunities for interaction; Interesting material	Personal interest; Look for ways to make personally relevant; Make it a game
Prior Knowledge Activation	Remind learner of prerequisite information, etc.	Ask self what is already known about the topic
Attention	Highlights; Animation; Audio supplements	Identify key features; Record notes
Encoding	Provide diagrams and/or multiple examples/contexts; Suggest mnemonics, etc.	Generate mneumonics, images and/or multiple examples/contexts
Comparison	Encourage comparison with diagrams/ charts/questions	Look for similarities; Draw diagrams/charts
Hypothesis Generation	Encourage student to think of and try various alternative courses of action	Generate possible alternatives and corresponding solutions
Repetition	Guided practice and/or reflection; Multiple perspectives/examples	Systematic reviews
Feedback	Provide instructively relevant feedback and correctives	Seek answers to self-posed questions
Evaluation	Have next action by student based on an evaluation of the feedback received	Ask "What do I currently know?" "What do I need to know?"
Monitoring	Check for understanding	Monitor Performance; Self testing
Combination, Integration, Synthesis	Provide ways to combine and integrate information - e.g. with graphics or multimedia	Establish categories; Construct tables; Seek higher-order relationships

grating advising functions into a learning environment, both aspects can be addressed by the same application.

In their discussion of teaching and learning principles in medical computer-based education (CBE), Jelovsek et. al (11) argue for the incorporation of established principles from educational theories. This will improve medical CBE applications considerably in acceptability and instructional efficiency. The necessity of basing CBE applications on established educational theories has further been shown by Lillehaug (13) in an evaluation of traditional CBE approaches within medical education.

Theories of instruction are not theories of learning (18). The critical element in this distinction lies in the description of means to achieve goals. Instructional theories are concerned with how actions taken by an instructional agent achieve instructional goals. In contrast, theories of learning are concerned with how actions taken by a learner achieve outcomes. However, the two are closely related. For instance, the formalisation of an instructional theory will often be influenced by the development of a corresponding learning theory.

Shuell (20) describes the relation between instruction and learning in his discussion of the student's role in learning from instruction. Shuell claims that learning is an active, constructive, cumulative and goal-oriented process where the learner is required to play an active part. For effective learning to occur, various psychological processes in the learner must be engaged. For example, learning must build upon, and be influenced by, the learner's prior knowledge, and certain functions integral to the learning process must be performed by either the learner or the instructional agent. Shuell has summarised current cognitive research on learning and education in a theory of twelve learning functions (see table 1). Each of these can be accomplished in a variety of effective and appropriate ways. Although the functions can be initiated by either the instructor or the learner, it is actually the learner who must carry them out. The functions are mechanisms to cue different learning processes. Lillehaug (13) argues that Shuell's theory of learning functions does not only summarise the current cognitive research on learning and education. It can also be used as a reminder scheme when designing learning environments or as a framework for evaluating the effectiveness of different educational settings such as computer-based education, through their support of the functions.

In the following section we will use Shuell's learning functions as a reminder scheme in proposing additional educational facilities for remote echocardiology in the form of computer-based applications. Also, it will be discussed how such applications can incorporate advisory facilities to help the practitioner.

4.1 Remote echocardiology and additional educational facilities

During remote echocardiology the specialist can trigger different learning functions as follows: focus the practitioner's *attention* on the colour and form that characterises regurgitated and stenotic valve lesions; encourage *comparison* by changing presentation between different modes of echocardiographic signals (e.g. from M-mode to doppler); make the practitioner *generate a hypothesis* in the form of a diagnosis (e.g. some cardiology dysfunction) whose investigation is given relevant instructional *feedback* and correctives (e.g. turn the probe 30 degrees to get a more informative view). During the entire session the specialist monitors the practitioner's performance (e.g. poses questions that check for understanding). The practitioner also has the opportunity to initiate the different learning functions himself (e.g. *comparison* by looking for similarities between a particular patient case and previous examined patients). A computer-based application for enhancing learning efficiency needs to support the learning functions in a similar way.

The first thing the practitioner has to learn is how the movements of the probe affect the presentation of echocardiology signals and when to make use of the different modes (good examination techniques). A learning environment can cover these aspects through integration of multimedia where the moves of the probe are shown in one picture frame while the corresponding output of echocardiology is shown in another frame. The movements of the probe can be accompanied by textual information which explains what is done and which effect this has on the output. By including possibilities to show outputs of all three modes at the same time, the relations between these can also be explained.

Typical patient cases with textual information could be stored (also by the specialist and the practitioner) in a database for reviews by the practitioner. Through options for freezing and focusing, the practitioner may freeze a case and focus on details of particular interest. Exclusion/inclusion of textual information can be used by the practitioner to test out his own hypothesis.

A learning environment with the described facilities may give support for Shuell's learning functions as follows:

- *Expectation and motivation:* Supported through overviews of what needs to be learned, and further, by pointers to the paths for how to learn it.
- *Prior knowledge activation:* In demonstrating techniques for how to use the probe, textual information/explanation reminds the practitioner of prerequisite general knowledge or particular information about the actual patient.
- *Attention:* Supported through visual and textual pointers to details of particular interest in a reviewed patient case. Further, the practitioner is given time to process all available information through speed/pause control. Also, a scratchpad editor might be usable to make notes and save information for future reference.
- *Encoding:* Supported by real-life patient cases examined with the same equipment as by the practitioner.
- *Comparison:* Supported through possibilities for comparing different typical cases stored in a database, or through comparing the practitioner's hypothesis with the corresponding cases in the database. Further, demonstrations of how to best make use of the different examination modes can be valuable.
- *Hypothesis generation:* Supported through possibilities for the practitioner to make hypothesis on patient cases which can be checked out versus cases in the database.
- *Repetition:* The practitioner can repeat the review of any case available in the database, or he can store own patient cases for later systematic reviews.
- *Feedback:* The practitioner can run a stored case, first without textual explanation, generate hypothesis, and get feedback through reviewing the case with textual explanations.
- *Evaluation:* Supported through feedback that can be applied by the practitioner in future cases.
- *Monitoring:* Supported by having the application check out the practitioner's suggestions for diagnosis and treatment plans on stored cases. In addition the practitioner can continuously monitor his own performance through the feedback he receives.
- *Combination, integration, synthesis:* Supported through a combination of demonstrations with and without explanations, comparisons between stored cases and practitioner's cases, and presentation of relations between the output of the different examination modes.

The suggested facilities for a computer application may support some of the practitioner's educational needs. Further, use of the application may make it possible for the practitioner to train/work without having the specialist available. Thus, the above discussions has implicitly indicated the potential for such an application to function as a kind of job aid (19). However, it is of utmost importance that the specialist still do supervising from time to time through ordinary remote consultations.

We have focused on, and discussed additional educational facilities for one particular remote consultation, namely remote echocardiology. It is a topic for further work to specify our initial suggestions in more detail, and to implement and investigate them. The possible generalisation of facilities for other kinds of remote consultations should also be put on the future agenda.

5 Conclusions

A number of services have been investigated and further developed within the NTR's telemedicine project (16,17). The use of remote medical consultations has not only been successful in the sense of making specialist medical services available to a larger part of the population, but also in enhancing the competence of the health care personnel in remote areas. Patients, practitioners and specialists are, in general, positive to the use of remote consultations.

The educational success obtained through remote consultations has been discussed and the learning effects have found support in theory of cognitive apprenticeship.

The current organisation of remote consultations has to be revised in order for the different hospitals to provide such services on a larger scale. The load on the specialists and their departments can be reduced by facilitating practitioners with supplementary computer-based applications. The specification of appropriate functionalities will benefit from consulting educational theory, e.g. from having Shuell's learning functions serve as a reminder scheme.

The consequences on society of offering specialist services based on remote consultations, where the practitioner can learn while performing his job, include among others that the stability, competence and recruiting of medical personnel in the districts may increase.

We have, in fact, discussed a service and an organisational set-up which facilitates on the job training, where learning occurs as you work, and may, thus, make the old vision of life-long learning a reality.

Acknowledgements

The authors thank Per Lunde, Steinar Pedersen, Svein Erik Stenvold for valuable discussions on their participation in the telemedicine projects, and Birger Nymo and Barbara Wasson for useful comments on earlier drafts of this paper.

References

1 Afseth, J E, Lunde, P. Distance education in echocardiology through videoconferences, *The Journal of the Norwegian Medical Association*. (In Norwegian, In Press).

2 AIM. *Supplement application of telecommunications for health care telemedicine.* (Advanced Informatics in Medicine. Commission of the European Communities, 1990) AI, 1685.

3 Akselsen, S, Eidsvik, A K, Folkow, T. Telemedicine and ISDN, *IEEE Communications Magazine*, 31(1), 46-51, 1993.

4 Brown, J S, Collins, A, Duguid, P. Situated cognition and the culture of learning, *Educational Research,* 18, 32-41, 1989.

5 Collins, A, Brown, J S. The computer as a tool for learning through reflection. In: H Mandl and A Lesgold (eds), *Learning issues for intelligent tutoring systems*. New York, Springer Verlag, 1988, 1-18.

6 Collins, A, Brown, J S, Newman, S E. Cognitive apprenticeship: Teaching students the crafts of reading, writing and mathematics. In L B Resnick (ed), *Knowing, learning and instruction: Essays in honour of Robert Glaser.* Hillsdale, NJ, Lawrence Erlbaum Associates, 1989, 455-494.

7 Hartviksen, G, Pedersen, S. Remote endoscopy of otorhinolaryngology patients. In: A Katzir (ed), *Optical fibres in medicine VII,* (SPIE Proceedings 1649), 1992.

8 Higgins, C, Dunn, E, Conrath, D. Telemedicine: a historical perspective, *Telecommunications Policy*, 1984.

9 Holand, U, Tetzchner, S von, Steindal, K. Two field trials with videophones in psychiatric and habilitative work. In: S von Tetzchner (ed), *Issues in telecommunication and disability.* COST 219, 1991.

10 Holand, U, Stenvold, S E. *Are the patients satisfied with teledermatological consultations?* (In Preparation).

11 Jelovsek, F R et al. Application of Teaching and Learning Principles to Computer-Aided Instruction. *M. D. Computing,* 6, 267-273, 1989.

12 Jøsendal, O et al. Distance diagnosis of skin diseases, *The Journal of the Norwegian Medical Association* 111, 20-22, 1991. (In Norwegian)

13 Lillehaug, S-I. ILE-VT: *Towards an intelligent learning environment in ventilator therapy.* Cand. Scient. Thesis. Department of computer science, University of Tromsø, Norway, 1992. (Also available as TF-education U6/92, Norwegian Telecom Research, Kjeller, 1992.)

14 Møllerbråten, B. Videotelephone development in Norway. In: T Kristiansen (ed), *A window to the future. The videotelephone experience in Norway*. Kjeller, Norwegian Telecom Research, 1991, 9-20.

15 Nordrum, I et al. Remote frozen section service: A telepathology project in Northern Norway, *Human Pathology* 22, 511-518, 1991.

16 Nymo, B (ed). Special issue on Telemedicine. *Telektronikk,* 1, 1993.

17 Nymo, B, Engum, B. *Telemedicine to improve the quality, availability and effectiveness of the health service in rural regions.* Kjeller, Norwegian Telecom Research, 1990. (TF lecture F10/90.)

18 Pirolli, P, Greeno, J G. The problem space of instructional design. In: J Psotka, L D Massey, S A Mutter (eds). *Intelligent tutoring systems: Lessons learned*. Hillsdale, NJ, Lawrence Erlbaum Associates, 1988.

19 Romiszowski, A. Expert systems in education and training: Automated job aids or sophisticated instructional media, *Educational Technology* 27, 22-30, 1987.

20 Shuell, T J. Designing instructional computing systems for meaningful learning. In: M Jones and P Winne (eds), *Foundations and frontiers of adaptive learning environments.* New York, Springer Verlag, 1992.

21 Sund, T, Rinde, E, Størmer, J. Full-scale replacement of a visiting radiologist service with teleradiology. In: H U Lemke et. al (eds), *Computer assisted radiology, Proceedings of the International Symposiums CAR-91.* Berlin, Springer-Verlag, 1991.

Telemedicine as a health-political means

BY STEINAR PEDERSEN AND UNNI HOLAND

621.39:61

One of the central health-political objectives of the Norwegian authorities is that the population should have equal access to medical services independent of their geographical location of residence.

Based on the knowledge of the geography, population density, access to general practitioners and medical specialists within the various medical fields, this paper will account for how we see telemedicine as a means to improve the utilisation of health resources. By this we will show how telemedicine can be a means of obtaining the objectives of the health authorities.

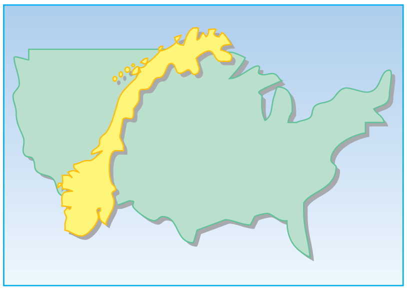


Figure 1

Introduction

Geography

Norway constitutes the western and northern part of the Scandinavian peninsula. The total area is 386,000 square kilometres (see figure 1).

Climatic conditions

Norway is situated in the northernmost part of the temperate zone. In extended periods of the year it has a rather varying climate.

Population

The country has 4.33 million inhabitants. The main part of the population lives in towns or densely built-up areas.

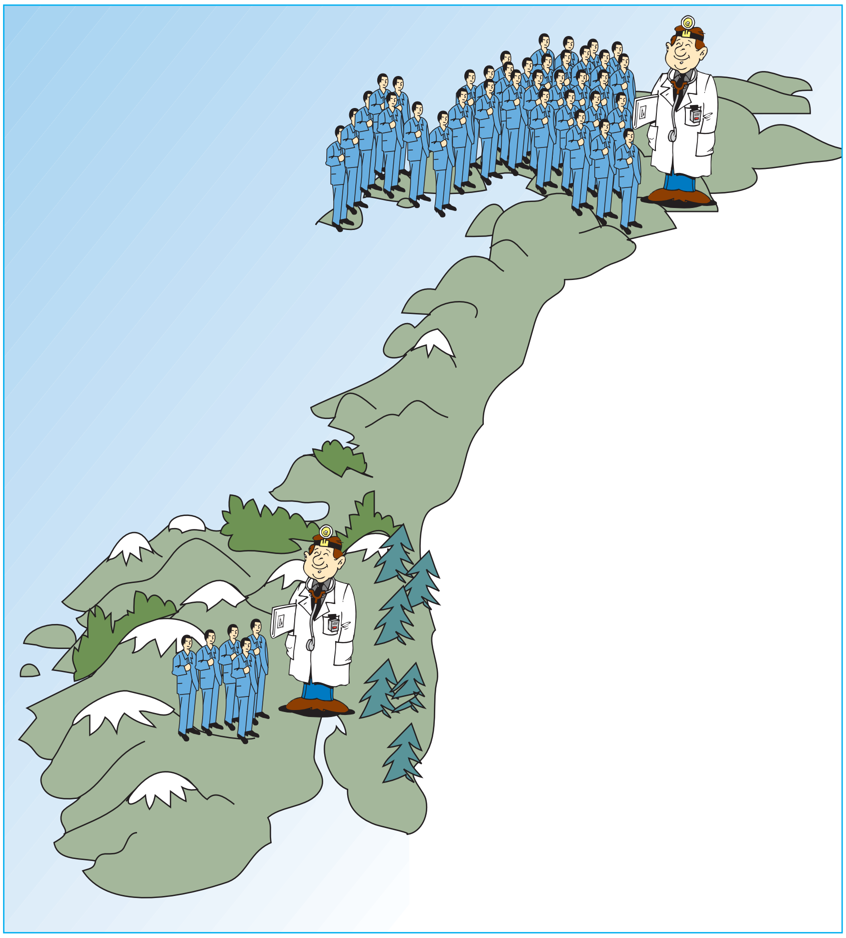


Figure 2

Health authorities

Administratively, Norway is divided into 20 counties. Each county consists of several municipalities or townships. The counties are responsible for the specialist health services. The municipalities are responsible for the primary health services. The Government funds framework contribution to the counties and the municipalities. Largely, the funds earmarked for the health service are related to the health care institutions, and as the various medical services are located in different geographical locations, many counties are thus dependent on buying health services they do not possess themselves.

Further, the country is divided into 5 health regions. Each region has a university clinic that, inter alia, is responsible for training of all medical practitioners.

Medical practitioners

Norwegian medical practitioners are trained during 6 1/2 year studies at the university. This practical and theoretical education is followed by one year with practical work at a hospital as well as half a year service at a county medical office. This is the prerequisite for authorisation for independent practise.

Of the 11,383 medical practitioners in the country, 5,880 has chosen to undergo 6 years additional training as medical specialists at a hospital. 565 specialists practise outside the hospitals. The majority of these are working in central eastern areas.

In Oslo, the capital of Norway, the ratio of patients to medical specialists is 291:1. In Finnmark, the northernmost county, the ratio is 1194:1 (see figure 2).

This rural-urban imbalance in the access to medical specialists reflects the increasing urbanisation which has taken place in all fields in Norway during the last 25 years. As far as doctors are concerned, the heavy work load and professional isolation in the rural areas are some of the main causes of centralisation of competence and specialised treatment.

Norwegian health authorities have – through various stimulation efforts – tried to promote more doctors to settle in rural areas. Such means as increased payment, longer holidays and subsidies for conference participation have been tried without any success worth mentioning. One of the main causes of the lacking success is that the efforts have not improved the doctors feeling of professional isolation (see figure 3).

Telemedicine

Telemedicine is defined as the investigation, monitoring and management of patients which allow ready access to expert advice and patient information no matter where the patient or relevant information is located.

Effect of telemedicine on the Norwegian public health service

Help to the patients and the medical practitioners

By help of telemedicine it is possible to give Norwegian general practitioners (GPs) immediate and direct access to medical experts in various hospitals. This reduces the feeling of professional isolation. As described above, this is one of the main causes of the poor coverage of doctors in rural areas.

In this way, telemedicine will have a twofold function by improving access to medical expertise, as well as at the same time stabilising the manpower situation among health workers in rural areas.

By help of telemedicine the rural general practitioners will – at least as long as it disseminated by video conference network – have an advantage in preference to their urban colleagues in being counsel for their patients by obtaining quick access to expert assistance.

Distribution of competence

We would like to summarise the main effects of new developments in telemedicine in the conception of building up decentralised competence. Besides giving the population in rural areas access to required medical expertise, it turns out that telemedicine contributes to a considerable professional strengthening of local GPs by means of the personal instruction that has been brought about.

Usually, remote consultation takes place in collaboration with the local doctor and the patient. This implicates that the local doctor can participate in the specialist examination. In turn, this leads to a greater insight in the medical problem. In the longer term the local doctors will by this means be able to treat far more patients who otherwise would have to be referred to outpatients department at the hospitals. In this way resources at the hospitals are released and travelling expenses are reduced.

Figure 3

The conditions for personal instruction can be further developed to distance education through the system of video conference. By reorganisation of the internal training at larger hospitals, it can be distributed via the telecommunication network to health workers at peripheral institutions.

For general practitioners, it is accepted that telemedicine counts for the speciality concerning skin diseases and psychiatry.

Help to better exploitation of resources

Telemedicine has no geographical boundaries. In principle, medical resources are available wherever the telecommunication network is accessible. The vision is that required medical expertise can be utilised at the place where it exists. In the future, it will not be essential that all medical centres possess all leading-edge competence in their midst. It can be sought where it can be found. In the near future the national health network can be extended with direct connection to professionals abroad.

Up to now, the selection of professional collaborators within the health network has to a great extent been regulated by economic frameworks. The future telemarket will be characterised by the fact that demand for medical competence will be directed towards centres that are distinguished by their medical competence, availability, price and general level of service.

Help to health-political planning

Telemedicine will also be an aid to better standardisation of treatment and control procedures for medical treatment plans. Where the telecommunication network is used to distribute information and advise, it could at the same time be considered a collection and adaptation of medical data about the population, standardised on a country basis. This would give medical researchers and health planners an improved foundation for future strategic decisions.

Critical factors

For the time being, technical solutions are not a limiting factor for development or utilisation of the potential of telemedicine. Today, good technical solutions are available in several fields. These will be further improved in the future. Furthermore, new technology will be developed, making more and more new services obtainable.

Figure 4

The cost of technical solutions and consequently, the use of them, will steadily decrease. The limiting factor will be the ability of the health authorities to adapt in order to fully use the telemedical potential.

In Norway the greatest obstacle will be the economic and administrative frameworks inherent in the fact that the Norwegian public health service is decentralised to 3 financing levels with geographical boundaries. As the telecommunication network does not take such boundaries into consideration, the present organising and administrative management structures must be altered and made more flexible. Otherwise, this will constitute a great threat to the present positive development. In addition, within the various administrations there will be persons who feel their own positions threatened. Furthermore, such groups of health personnel who have built up their positions by being the sole holders of medical competence will have their positions threatened and may act as saboteurs.

Department of telemedicine

In order to consolidate the telemedical competence in Norway the Ministry of Health and Social Affairs has assigned the nationwide function for telemedicine to the University Hospital of Tromsø. A special department is now being established through co-operation between researchers at the Norwegian Telecom Research in Tromsø and medical personnel at the University Hospital of Tromsø. In addition to developing new applications for telemedicine, the Department of Telemedicine shall guarantee that tele-medical services match medical standards. It shall also further develop legal and organisational conditions within telemedicine, propose new payment arrangements for the service and seek to put approved applications into operation on a nationwide basis.

In its preliminary work this telemedical department has proposed towards the central health authorities that the legal responsibilities towards the individual patient in a telemedical consultation situation shall not differ from any other patient/doctor situation. This means that it is always the person with highest medical competence who decides whether the patient information upon which a decision is made, is of satisfactory quality. This principle must be considered independent of the way in which the information is obtained by the specialist.

Responsibility to prevent leakage of data in the telecommunication network is imposed on the company delivering the service. Responsibility for storage of data at health institutions must be imposed on the individual institution in accordance with further regulations by superior data supervision and health authorities. The telemedical department has further proposed that payment for telemedical services should be channelled to the department which has rendered the service. It is our opinion that the economic profit inherent in the reorganisation into a medical nationwide health network forms a good basis for such a financing model. The central health authorities would then be the financing sources.

When the premises for infrastructure within telecommunication are present and the required reorganisation of the public health services is complete, telemedicine will be the potential means needed by the health authorities to reach the objective of giving Norway's population equal access to medical services independent of their geographic location of residence.

Quality requirements for telemedical services

BY UNNI HOLAND AND STEINAR PEDERSEN

621.39:61
658.56

Telemedicine implies the use of new methods, routines and aids within the public health services. The telemedical history is short and experience is limited. Today many health workers and patients meet the telemedical services for the first time.

Among others, this is one reason why the following question is of central importance: Can patients and health workers have the same confidence in a diagnosis when the examination is carried out through telecommunication as when it is carried out in a traditional way? This question concerns the quality of the telemedical services. Justification and diffusion into the health services presuppose a positive answer to this and related questions.

Quality assurance is steadily gaining importance within the health services. Diagnostic confidence is a central point in the telemedical work. In the course of time the criteria for quality have changed somewhat. Only a few years ago quality would be assessed on the basis of criteria like equal offer of health services and fairness in the access to medical services. Today we see how the patients' own assessment and experience of the service are taken into account, coupled with the fact that quality tends to be assessed based on efficiency, productivity and other factors of practical economics. Quality cannot be assessed isolated from the ideological and health-political priorities applied at a certain time. To strive for the best possible quality entails a balance of various values.

Examination of service quality assumes an adequate base for comparison. To what standards of quality should telemedicine be compared? It is irrelevant to set up a standard requiring 100 % correct diagnosis in all cases, unquestionable satisfaction by the patients, or increased efficiency of all tasks encountered. Such an absolute standard is not applied by assessment of other functions of the health services either.

Assessment of quality should in our opinion be undertaken by a comparison of relative standard, based upon established practice by the health service. As an example, we know that the diagnostic practice implies a certain probability of mistakes. Such errors emerge either by diagnosis of a disease not present in the patient (false positives) or by a complaint not detected and diagnosed (false negatives).

Within the health services there seem to exist unwritten and fraternal general lines for acceptable levels of erroneous diagnosis. Such levels will vary depending on the field and environment in the health service. In spite of these variations, it is these quality requirements that telemedicine has to be put up against.

One recurring problem is the fact that documentation of the qualitative level of today's examination methods is often lacking. Examinations of quality of telemedical services have brought to light the lack of corresponding data by use of traditional methods.

Within telemedicine we will argue that quality should be discussed based upon two main conditions: first, an objective measure on performance of the medical practice and the management of competence and resources; and second, patient satisfaction concerning the telemedical services.

Standard of exercising medical service

The first main criterion for quality is related to medical and professional problems and the practising of medical competence. In concrete terms this implies:

- diagnostic security
- whether the help is given in adequate place and at adequate time
- whether the way of working adds to the competence and skill by the local health services in the long run.

Telemedical services must satisfy the demand on precise medical assessments and diagnosing. This is termed diagnostic confidence and constitutes an essential condition for application and further development of telemedicine.

Diagnosing has a central function within the health services. In the first place the diagnosis gives directions for treatment of the individual patient. In broad outline, the same diagnosis for several patients implies that they shall have equal treatment.

Another central function of the diagnosis is as "label" on patients with nearly common medical status. Such "labels" make it possible to exchange information about the phenomenon indicated by the diagnosis. In this way it is possible to systematise experience gained by the health workers, draw up strategies for the running of the institution and set up priorities in accordance with the population's condition of health.

A requirement to telemedical services must be that the probability of erroneous diagnoses does not exceed the error of traditional procedures. In particular, this requirement must cover the danger of false negatives.

The task of assuring diagnostic quality is an important part of the testing of the telemedical services. This is accomplished by a systematic comparison of the diagnostic precision of telemedical and traditional diagnosing respectively. Results from such investigations are available within the fields of dermatology (Jøsendal and Fosse, 1991), pathology (Eide, Nordrum and Stalsberg, 1992) and endoscopy (Hartviksen and Pedersen, 1992). All of these papers verify that telebased diagnosing maintains a high professional standard. In addition to these investigations, work is carried out on new quality studies in connection with the use of endoscopy and investigations are initiated in the field of cardiology. Preliminary results from the ongoing studies indicates that remote consultations within cardiology also satisfies the stringent requirements for diagnostic confidence.

Through remote diagnosing or by obtaining external competence in the work of a local practitioner, telemedicine makes it possible to deliver health services at the right point of time and at the right place. The patient can be examined, diagnosed, and have treatment as soon as it is appropriate from a professional assessment.

This effect has emerged in spite of the short period of time the telemedical services have been operational. The use of telemedicine leads to a genuine reduction of waiting time between reference and examination by a specialist. Early identification of the problems and consequent treatment improves the probability of effective recovery. This reduces the need for further referral for the same patient, which in turn leads to reduced waiting time for new patients. Telemedical services also provides access to medical expertise on very short notice, e.g. in an acute situation. Reduced travel activity increases the efficiency of the doctor.

Reduced waiting time and more efficient use of resources are related to the fact that telemedicine reduces the relevance of geographic address. In principle, it makes no difference whether the distance between doctor and patient is short or

Figure 1

long. Telemedicine makes possible a decentralisation of health services so that the patient to a larger extent can be treated near to his home. Medical examinations which used to consume a lot of time and money for travel activity either by the patient or by the medical specialist, can today be carried out through telemedicine.

Today we have no investigations which can give reliable answers to the extent of this effect. However, following extensive experiments and operation the experience is unambiguous. Health personnel give account of significant reduction in waiting time for the patients and travel activity for the doctors. Assessed by criteria for delivery of health services at an adequate point of time and as near the residence of the patient as possible, we claim that telemedicine maintains a higher level of quality than traditional medical services. As telemedicine in practice constitutes a supplement to traditional methods, telemedicine contributes to a general raise of quality of the health services in relation to those criteria.

The quality of a service should also be assessed in relation to its effect on the local conditions. To what extent can telemedicine contribute in making the local health services self-supported through enhancing competence and skill?

Reports from local health workers (Jøsendal and Fosse, 1991, Henriksen and Pedersen 1992, Hoff 1992, Haga 1993) show that telemedicine enhances their ability to master medical challenges, so that they gradually can handle problems that earlier required reference to specialists.

In practice remote consultation implies personal instruction to the GP. He or she is present during the consultation together with the patient and takes part in the examination by the specialist. This has a great and immediate effect on the doctor's professional level and competence. If we here, as in other medical connections, assume that GPs and health workers have a critical relation to their own competence, such an effect of telemedicine is solely positive and satisfy the demands for quality.

Patient satisfaction

Another criterion of quality is the patients' own satisfaction with remote diagnosing. Among other things, patient satisfaction will be determined by whether they experience and believe that the service is of high medical quality, and whether they are satisfied by the way the remote consultation is performed.

Patients' satisfaction with the health service seems to become an increasingly more important argument in the health-political discussions. Administrators and politicians regard patients as customers and the product delivered must satisfy the customers' needs and expectations. In this way a high degree of patient satisfaction becomes a goal in itself and constitutes one of several criteria for service quality.

Lack of patient satisfaction may also have practical consequences of a negative nature. Investigations (re. Wolf et al., 1978) shows that unsatisfied patients to a greater degree than satisfied patients neglect to follow instructions and prescriptions by the doctor. We assume that such behaviour reduces the probability for quick recovery. Satisfied patients are less inclined to go "doctor shopping" (re. Pascoe, 1983). Further, patient satisfaction seems to have a bearing on the their inherent self-healing forces (Wolf et al, 1978). Satisfied patients use their own resources towards the recovery process to a higher degree than unsatisfied patients. Such conditions turn patient satisfaction into more than a question of the patient's personal experience of the treatment. It also implies consequences for health behaviour and exploitation of resources.

Investigations have been initiated on satisfaction among patients who have been assessed or diagnosed by help of different telemedical methods. Investigations are in progress to examine patient satisfaction after treatment by tele-endoscopy, teledermatology and telepsychiatry.

Before the ongoing inquiries are completed we must be reserved in drawing conclusions. As far as available material is analysed, the impression is that the patients are well satisfied by remote diagnosis. This has emerged from questionnaires and interviews of patients.

Investigations of patients' satisfaction with remote diagnosis of skin diseases show in broad outline the following facts (Holand & Stenvold, 1993):

Close to half of the patients suspect that they would have received better help for their ailments if the dermatologist had come to the local hospital and carried out the examination in person. In spite of this doubt whether telebased examinations give optimal medical help, 66 % of the patients express their satisfaction with the telemedical examination, close to 20 % were neither satisfied nor dissatisfied, 12 % were dissatisfied while 4 % had no opinion.

Some advantages emphasized by these investigations are reduced waiting time, saving of resources, reliability of resources because both doctors (directly or indirectly) took part and positive excitement by application of new technology. Possible disadvantages that have emerged point to lack of personal contact, uneasiness about being filmed, and a certain doubt about the thoroughness of the examination due to lack of skin contact. 44 % of the patients say they would prefer a new telemedical consultation in case they should need a renewed specialist examination. 30 % would prefer the specialist to come to their local hospital. 18 % would prefer to go to the regional hospital for examination.

The quoted and corresponding data form a basis for great optimism on behalf of telemedicine concerning the patients' own assessment of quality and to what degree they are satisfied with this method of examination.

Conclusion

Much work remains concerning quality aspects of telemedicine. First, relevant quality requirements to the various telemedical fields must be put into operation. Second, controls must be carried out to assure that the services are in line with these requirements. Third, routines must be established to assure the maintenance of the level of quality and possible improvement of the quality.

As telemedicine gradually capture a more natural place in the activities of hospitals and the primary health service, it seems reasonable that they should have the main responsibility for quality requirements and quality control. An optimum solution would be for quality to be continuously assessed on the basis of medical requirements, technical requirements, and patients' own requirements.

References

Eide, T J, Nordrum, I, Stalsberg, H. 1992. The validity of frozen section diagnosis based on video-microscopy. In press, *Zentralblatt für Patologie.*

Haga, D. 1993. Opinion presented in *Sosialkanalen,* Norwegian Broadcasting System, TV, 13.01.93.

Hartviksen, G, Pedersen, S. 1992. Remote endoscopy of otorhinolaryngology patients. In: *Proceedings of Optical Fibres in Medicine VII,* ed. by A Katzir, Proceedings SPIE; vol 1649, 186-91.

Hoff, I. 1992. Video: *Telepsychiatry. Another way of meeting.* Kåre Tannvik A/S, Kirkenes. (In Norwegian)

Holand, U, Stenvold, S E. 1993. *Are the patients satisfied with remote teledermatological consultations?* In prep. for First international conference on medical aspects of Telemedicine, Tromsø.

Jøsendal, O et al. 1991. Remote diagnosis of skin complaints. *The Journal of the Norwegian Medical Association,* 3(111), 343-5. (In Norwegian)

Pascoe, G C. 1991. Patient satisfaction in primary health care. *Evaluation and program planning,* 6, 185-210.

Wolf, M H et al. 1978. The medical interview satisfaction scale: Development of a scale to measure patient perception of physician behaviour. *Journal of Behavioural Medicine,* 1.

Standards for health care telematics – A new dimension and challenge to standards makers

BY SIGURD FROM

621.39:61
006

1 Introduction

The implementation of telemedicine applications on a large scale requires standards for the exchange of health care information to be in place. Norwegian Telecom participates in several standardisation projects both internationally and nationally, which develop standards for electronic exchange of health care information. This paper presents the work related to CEN/TC 251, the European initiative for establishing standards in health care informatics and telematics. The objective of this paper is to introduce a method for structuring health care telematics standards.

2 Standardisation

2.1 The traditional motivation for developing standards

The development of standards has traditionally played an important role in the spread of technology in Western society. In the highly specialised industry, products are assembled into new more complex products. A producer of a product may require many suppliers for the component products. Each component must have a well defined functionality. In order to achieve an acceptable price for the component, the same component should be usable in many products. The many-to-many relationship between producers and suppliers requires open specification - standards - to be handled efficiently. This is illustrated in figure 1.

2.2 ISO's definition of a standard

In order to establish a standard, the interested parties come together and form standard committees. ISO - the International Standardisation Organisation - is the largest standard making body in the world and holds a number of technical committees which develop international standards in many fields. ISO defines a standard to be:

Document established by consensus and approved by a recognised body, that provides, for common and repeated use, rules, guidelines or characteristics for activities or their results aimed at the achievement of a maximum degree in a given context.

Standards should be based on the consolidated results of science, technology and experience, and aimed at the promotion of optimum community benefits.

In practice, standards are documents that are used to set requirements to products so that the buyer has a guarantee for the quality of the product. To verify that a product conforms to standards, third party test laboratories, which certificate the products, have been set up.

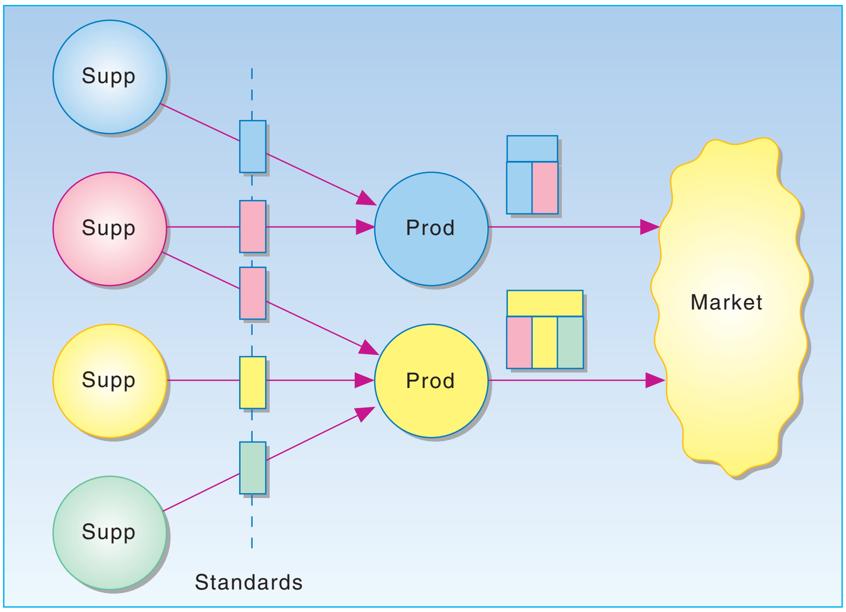


Figure 1 Many-to-many relationship between producers and suppliers require standards

In later years, the EC and EFTA countries have, as a result of preparing the common European market, increased the standardisation and harmonisation activity. Standards are used instead of governmental regulations. The standardisation process within the European standardisation bodies, CEN, CENELEC and ETSI, is open for participation from all interested parties in Europe. Experts in all fields are encouraged to participate. The final formal votes are made by delegates from the national standardisation institutes.

3 Health care telematics standards

3.1 The need for standards in health care telematics

As indicated both in other papers in this journal (From, 1993) and elsewhere (De Moor, 1993) a number of telemedicine and other health care telematics applications are under development.

Health care telematics applications require electronic exchange of information between computer systems. In order to fulfil this requirement, the systems need to share health care concepts. Only those concepts which are supported by both the sending and receiving system can be interpreted correctly and result in meaningful information exchange. This is illustrated in figure 2.

Information exchanged between systems must be represented technically in such a way that it can be transported across electronic networks. Today, a number of health care applications emerge, using their own interchange formats for health care messages. This makes it difficult and expensive to integrate systems from different vendors.

Standards in health care telematics are needed to define both the health care requirements such as the health care concepts, and the technical solutions such as interchange formats.

3.2 Categories of standards

Health care telematics can be described along two dimensions:

1 The health care dimension which includes health care subjects, health care services and their organisation.

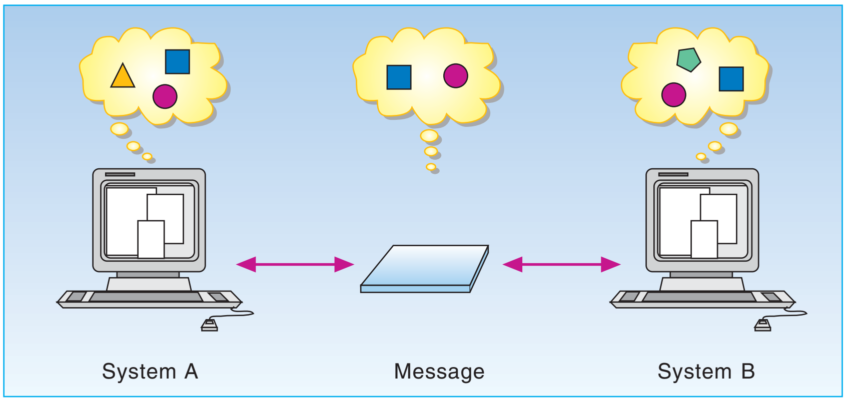


Figure 2 Only shared concepts can be exchanged in messages

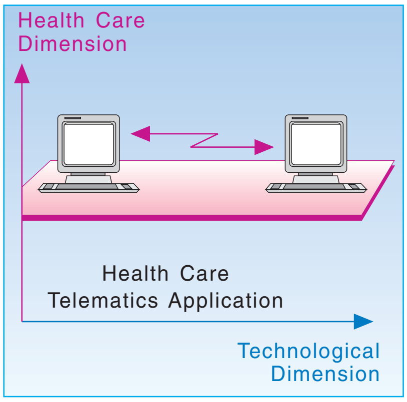


Figure 3 The health care and technological dimensions of health care telematics applications

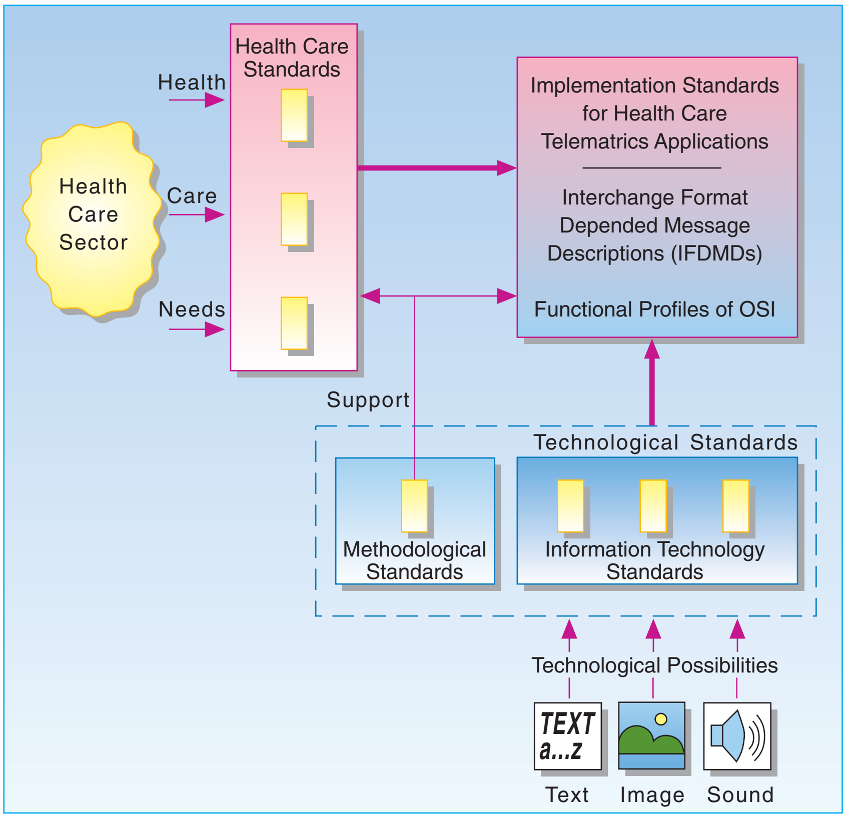


Figure 4 The relationships between categories of standards

2 The technological dimension which includes all supporting technology to realise the health care telematics applications.

The two dimensions are illustrated in figure 3.

Normally health care telematics standards will include aspects of both dimensions. However, a number of technological standards have already been established and are in common use in other part of society. These standards, such as the OSI-standards for communication, should be part of the technological platform for realising health care telematics applications. This must be considered in the standardisation process for health care telematics.

It may be useful to distinguish between different types of standards depending on whether their primary focus are health care or technology. For the remainder of this paper four categories of standards will be used:

1 Health care standards (HC-standards)

2 Information technology standards (IT-standards)

3 Implementation standards of health care telematics (Im-standards)

4 Supporting and Methodological standards (SM-standards)

All of these standards except category 2, IT-standards, may be called health care telematics standards. Figure 4 shows the relationship between the various categories of standards.

The *health care standards (HC-standards)* describe common health care needs. Today, manual paper forms for information exchange are created locally without co-ordination with other health care organisations. The lack of co-ordination results in a diversity in the use of terminology for health care concepts. This diversity becomes a problem when the paper forms are substituted by electronic communication systems. Common health care requirements are needed to allow for information exchange between heterogeneous systems and to reduce production and maintenance cost of the systems. This category covers standards such as Health Care Domain Descriptions (HCDDs), Communication Scenario Descriptions (CSDs), Coding systems, and Security Requirements.

The second category is the *Information Technology Standards (IT-standards).* This category includes standards for open systems interconnections (OSI), system architectures, user interfaces, storage, etc., which are independent of health care.

In order to realise a health care telematics application, HC-standards and IT-standards must be combined in an implementation. As shown in figure 4, *Implementation standards (Im-standards)* for health care telematics applications are created by mapping HC-standards onto IT-standards. This mapping process is

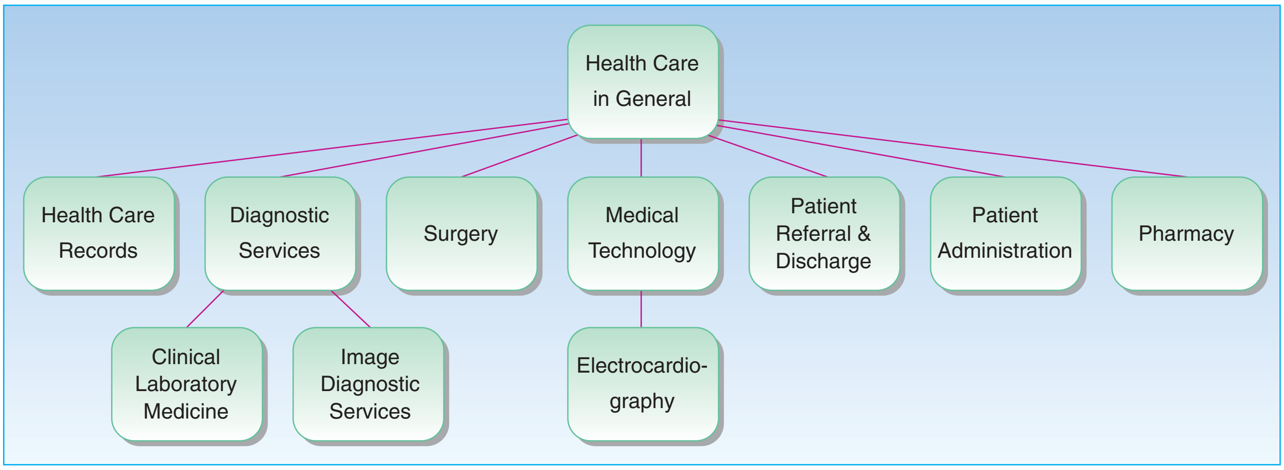


Figure 5 Health care domains and their relationships

guided by special mapping standards defined in category 4. IM-standards cover Interchange Format Dependent Message Descriptions (IFDMDs), Functional Profiles of OSI-standards for use in health care, and different types of Implementation Guidelines.

The *Supporting and Methodological Standards (SM-standards)* are technological standards developed specifically to support the establishment of health care telematics standards. They are not themselves used in a health care telematics applications. This category includes methods for developing HCDDs and CSDs and a metamodel for coding systems. In addition supporting standards such as registration procedure for coding systems and medical informatics vocabulary are included.

3.3 CEN TC/251 on standards in health care informatics

The European Committee for Standardisation (CEN) established in 1990 a technical committee for health care informatics, CEN/TC 251. The objectives of CEN/TC 251 are the organisation, the co-ordination and the follow-up of standards development, including testing standards in health care informatics and telematics at a European level (12 EC-countries, 7 EFTA-countries, 7 Eastern European countries). The development of such standards requires people with both medical and technical skills.

The final standards from CEN/TC 251 are aimed for use in the health care sector. CEN/TC 251 therefore only develops HC-standards, Im-standards and SM-standards. IT-standards are imported from other standards groups such as ISO/IEC JTC1, CCITT, ETSI, and EWOS.

CEN/TC 251 has defined a work programme (CEN, 1993). The work is divided into 7 working groups:

- WG 1 Health Care Information Modelling and Medical Records
- WG 2 Health Care Terminology, Semantics and Knowledge Bases
- WG 3 Health Care Communications and Messages
- WG 4 Medical Imaging and Multimedia
- WG 5 Medical Devices (communication with)
- WG 6 Health Care Security and Privacy, Quality and Safety
- WG 7 Intermittently Connected Devices (incl. patient-data-cards).

Each working group supervises a number of Project Teams (PTs). The work in a PT is undertaken by especially assigned experts and is funded. Until April 1993 the following PTs have been established:

Project Team	Monitoring WG	Description
PT 001	WG 1	Medical informatics vocabulary
PT 002	WG 2	Terminology and coding systems of medical procedures
PT 003	WG 2	Model for representation of semantics
PT 004	WG 3	Investigation of syntaxes for existing Interchange Formats to be used in health care
PT 005	WG 3	Procedures for registration of coding systems related to health care
PT 006	WG 4	Medical image and related data information format standards
PT 007	WG 5	Standard interchange format and communication protocol for computerised electrocardiography
PT 008	WG 3	Messages for exchange of laboratory information
PT 009	WG 7	Identification, administrative and common clinical data content for intermittently connected devices used in health care
PT 010	WG 1	Health care information framework
PT 011	WG 1	Electronic health care records architecture
PT 012	WG 6	Security for health care information systems
PT 024 (EWOS)	WG 3	Functional profiles for medical image interchange.

In addition to WGs and PTs, CEN/TC 251 has close liaisons with many other European and International organisations. Two of these are the EWOS/EG MED (European Workshop on Open Systems, Expert Group Medical) which handles functional profiles of OSI for use in health care, and the WEEB MD9 (Western European EDIFACT Board, Message Development Group for Health Care) which develops EDIFACT messages for health care which fall in the category of Im-standards.

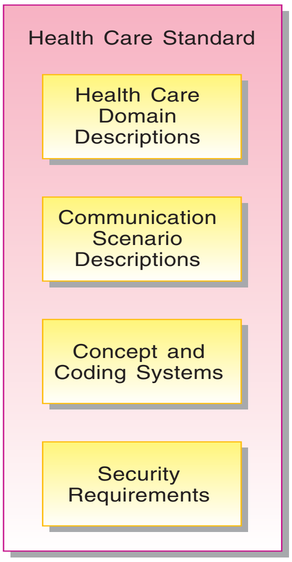


Figure 6 Categories of health care standards

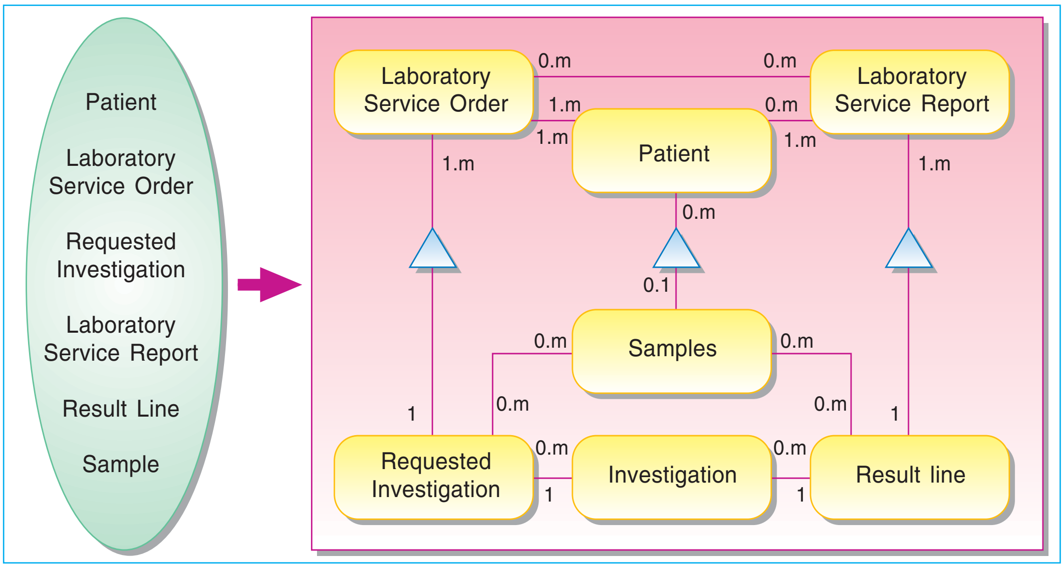


Figure 7 Concepts and domain information model taken from the clinical laboratory medicine domain

4 Health care standards

HC-standards describe solutions to common needs in a health care sector. HC-standards should be the result of a consensus process involving health care experts.

4.1 Health care domains

The health care sector may be separated into health care domains. CEN/TC 251 is addressing a number of domain in its work programme. Figure 5 indicates the health care domains currently under study and put them into a hierarchical structure to show their dependencies. The most general domain is found at the top; the most specific domains at the bottom.

The figure shows that clinical laboratory medicine is a sub-domain of the diagnostic services domain which means that clinical laboratory medicine is a specialisation of diagnostic services. Similarly, the diagnostic service domain is a sub-domain of health care in general. The figure only shows the specialisation-generalisation structure between the domains. Other relationships also exist but are not discussed here.

4.2 Categories of health care standards

Each health care domain has different properties which are described by different types of standards. We have chosen to categorise the health care standards into four groups as shown in figure 6.

4.2.1 Health care domain descriptions (HCDDs)

For each health care domain addressed, the user requirements need to be formally documented. The Health Care Domain Description (HCDD) defines the concepts used in a domain. As an example for the laboratory medicine domain, concepts such as laboratory service order, patient, investigation requested, are defined in the HCDD. In addition to the concept definitions, the HCDD includes a Domain Information Model (DIM) which defines the relationships between each concept in the domain. Figure 7 shows an example of a list of concepts and a DIM.

4.2.2 Communication scenario descriptions (CSDs)

In Open-Edi, a scenario is defined as a formal description of a class of business activities (Open-Edi, 1991). In health care the *Communication Scenario Descriptions* (CSDs) are used to describe information exchange between *communication parties*. A communication party is a health care person or organisation which participate in the communication. Each communication party has one or more *communication role* which defines their behaviour in a communication situation. For instance, a doctor requesting a laboratory analysis is a requester. It is the role as requester that is important in a laboratory requesting situation. Similarly the laboratory is the provider. Communication roles may be held by different communication parties. In the example, the doctor as a requester may be substituted by a nurse or a veterinarian who issues a laboratory request.

A set of *services* are associated with each communication role. A service defines a specific behaviour that the party holding the role is responsible for exhibiting when the service is requested.

The CSD shows the logical messages exchanged between the communication roles. A simplified communication scenario taken from the Laboratory Medicine Domain is shown in figure 8.

A logical message incorporates a selected amount of information which is exchanged between two systems for a given purpose e.g. a request for a service or a report from a performed service. The General Message Descriptions (GMDs) are defined as the result of combining a HCDD with a CSD. This is shown in figure 9.

For each logical message in the CSD a view of the Domain Information Model (DIM) in the HCDD is created. This becomes the GMD. One HCDD may relate to many CSDs. Similarly, one CSD may relate to many HCDDs. For each instance of the relationship between HCDDs and CSDs, one or more GMDs are created.

A GMD resulting in the combination of the logical message "New Laboratory Service Order" in the scenario above and the DIM described in figure 7 is shown in figure 10.

4.2.3 Concepts and coding systems

Most health care information processing requires the use of medical coding sys-

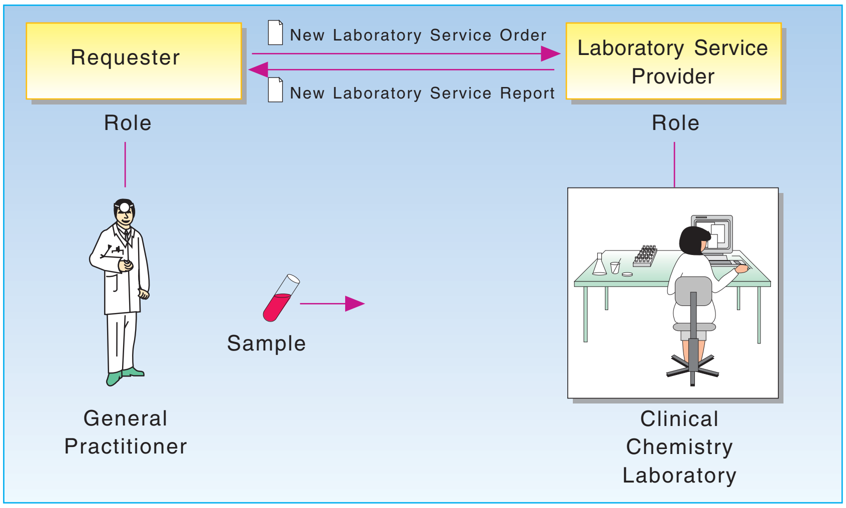


Figure 8 One simplified CSD related to the clinical laboratory medicine domain

tems. Coding systems are often domain dependent. The preference for coding systems varies between the users and there is often more than one system used in a domain. We have chosen to handle coding systems as a separate type of health care standards.

4.2.4 Security requirements

Security is an important aspect which influences almost all areas of health care telematics. Security includes issues such as confidentiality, integrity and availability. It is important to clarify and harmonise the security requirements for each health care domain and for health care in general.

4.3 Relationship to CEN/TC 251's Work Programme

The development of HC-standards are undertaken by the different WGs in CEN/TC 251. Each WG has been assigned a number of work items described in the CEN/TC 251 Work Program. Table 1 shows the relationship between HC domains, WGs and work items. For each work item the type of HC-standard is indicated by the colour of the intersection.

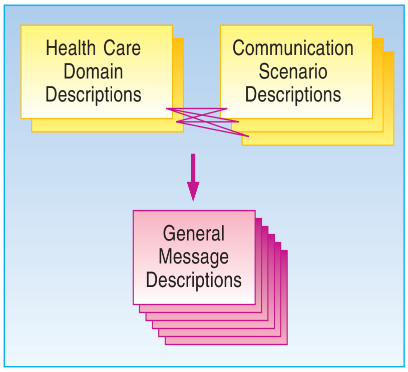


Figure 9 Relationships between HCDDs, CSDs, and GMDs

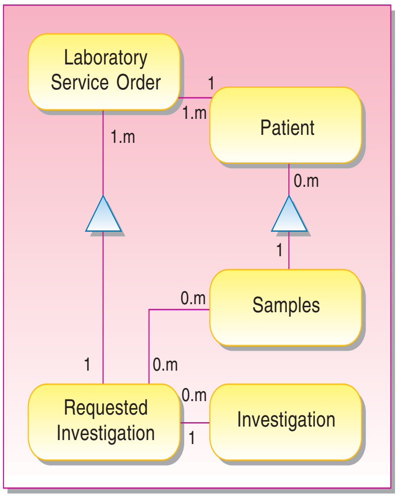


Figure 10 GMD for new laboratory service order

Table 1 Work items in the CEN/TC 251 Work Program sorted on HC domains

No	Domain	WG 1	WG 2	WG 3	WG 4	WG 5	WG 6	WG 7
1	Health Care in General	WI 1.1 PT 010 HCI-Frame					WI 6.2 PT 012 COMPUSEC	
2	Health Care Records	WI 1.6 PT 011 EHCR-A						
3	Diagnostic Services			WI 3.10 DIAMES				
3.1	Clinical Laboratory Medicine		WI 2.4 PT 002 PROCTERM	WI 3.5 PT 008 LABMES				
3.2	Image Diagnostic Services				WI 4.3 PT 006 MIF			
4	Surgery		WI 2.4 PT 002 PROCTERM					
5	Medical Technology							WI 7.0 PT 009 OLDIF
5.1	Electro-cardiography					WI 5.2 PT 007 SCE		
6	Patient Referral and Discharge			WI 3.12 MPRD				
7	Patient Administration			WI 3.11 ADMES				
8	Pharmacy			WI 3.15 DRUGSPEC				

HCDD and CSD — Coding System — Security Requirement

Work item 1.6: The health care record is an extended version of the medical record which covers all types of health care information related to a patient. The objective of this work item is to define an overall architecture for the health care record and the various components of it.

5 Information technology standards

CEN/TC 251's objective is to develop standards in the field of health care informatics and telematics. It is not the intention to develop new standards for new technology. Technology standards, and particular IT-standards, have their origin in bodies such as ISO/IEC JTC 1, CCITT, ETSI and EWOS. CEN/TC 251 should import IT-standards from these international and European bodies. The number of IT-standards is large. The categories most relevant to health care telematics applications are:

- Interchange formats (ASN.1, EDI-FACT, IPI, etc.)
- Character repertoire (ASCII, ISO 8859, ISO 10646, etc.)
- OSI Application standards (FTAM, RDA, TP, X.400, X.500, etc.)
- OSI Transport standards (ISDN, X.25, CSMA/CD, Token ring, FDDI, etc.)
- Security standards (ISO 7498-2, ISO 10181, etc.).

The categories are illustrated in figure 11.

Standards for bar-codes is an example of another type of technological standard that may be required in e.g. communication between general practitioners and laboratories in order to identify samples unambiguously.

6 Implementation standards of health care telematics

Im-standards are created by mapping HC-standards onto IT-standards guided by special mapping rules. Figure 12 shows three categories of Im-standards: Interchange Format Depended Message Descriptions (IFDMD), Functional Profiles of OSI, and Implementation Guidelines.

6.1 Interchange format depended message descriptions

IFDMDs are created by mapping GMDs to a selected Interchange Format (IF). One GMD may be mapped to different

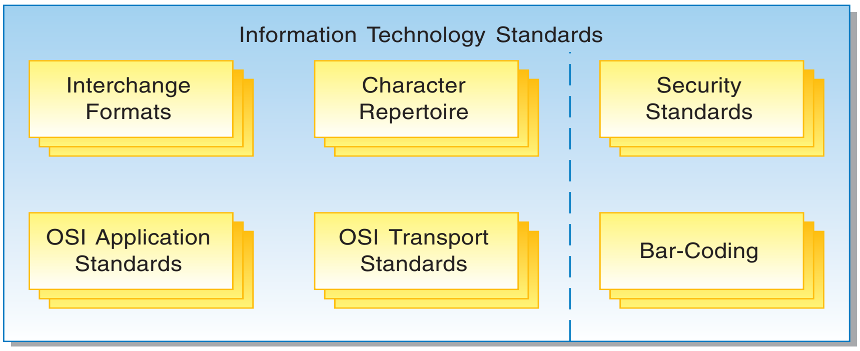


Figure 11 Information technology standards

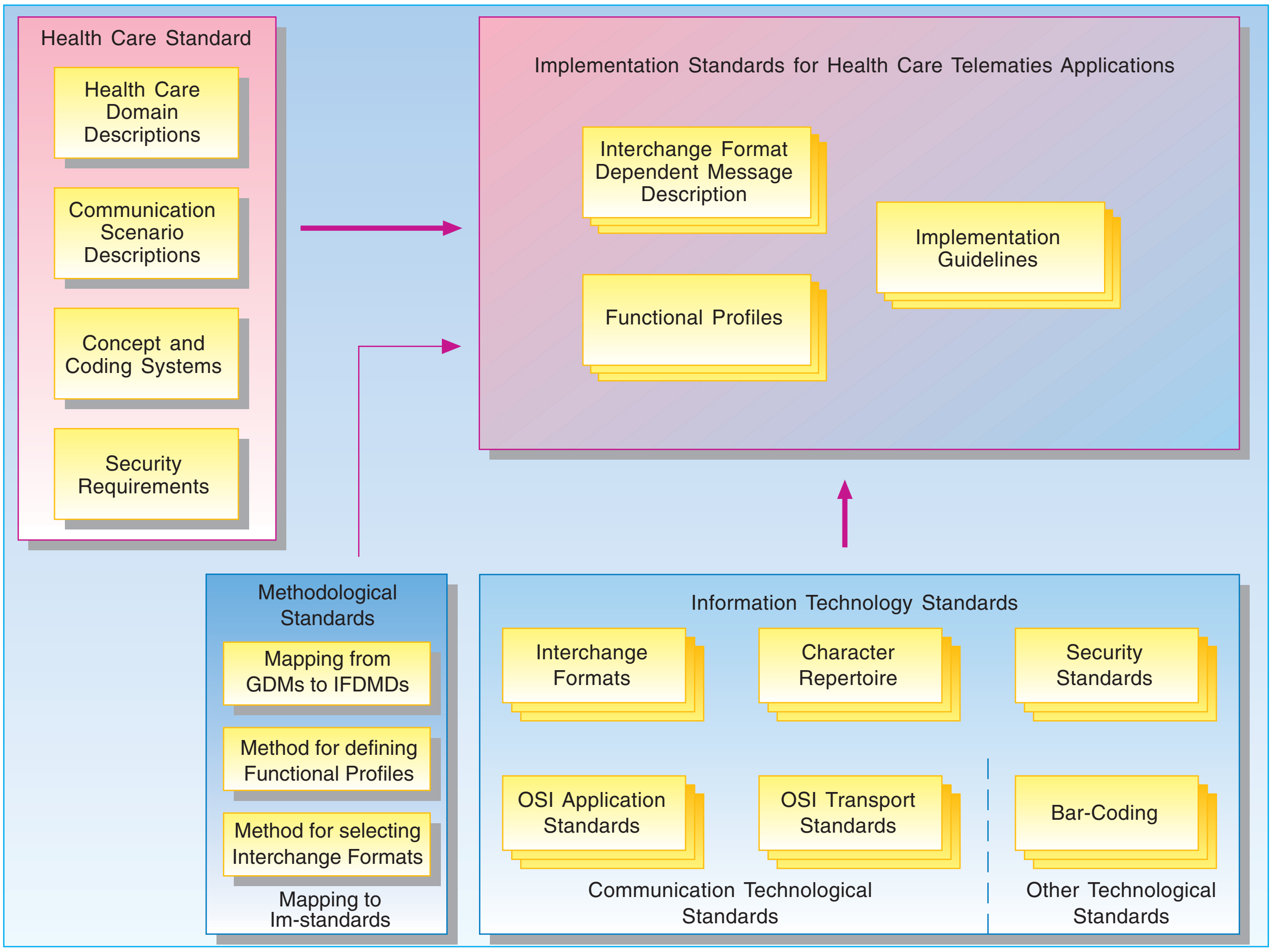


Figure 12 Implementation standards

IFDMD using different IFs. This is shown in figure 13.

The first step in a mapping should be to select the IF to use. The requirements to information structure and data types given by the GMD and the different IFs' ability to meet these requirements, should determine which IF is selected. PT 004 has developed a method for selecting interchange formats based on GMD requirements which fall into the category of SM-standards.

The second step is the actual mapping where objects and attributes in the GMD is mapped to the data elements of the IFDMD. The rules for this mapping will depend on the IF selected. PT 008 has as a part of the work of mapping GMDs to EDIFACT messages defined a preliminary method for this mapping.

6.2 Functional profiles

Functional profiles are needed to ensure communication between open systems. Many options exist in the OSI base standards, which have to be resolved before real interoperability can be achieved. This is the objective of Functional Profiles. A Functional Profile may include one or more OSI base standards. Examples of Functional Profiles are found in the national OSI profiles, e.g. GOSIP and NOSIP.

Most of the health care telematics applications should use existing Functional Profiles defined elsewhere. However, some applications will require special profiles. One such area is medical imaging, where large amounts of data need to be transferred fast across the network.

EWOS EG MED has defined a method whereby "real world" user requirements for communication between health care systems can be mapped on to open systems profiles. The method may be used both to select existing OSI Functional Profiles and to indicate where there are needs for developing new Functional Profiles (EWOS, 1992).

6.3 Implementation guidelines

The functional requirements to health care telematics applications may vary from country to country. Still, there may be a need to ensure standard solutions within a country. Implementation Guidelines may be used to add additional implementation requirements.

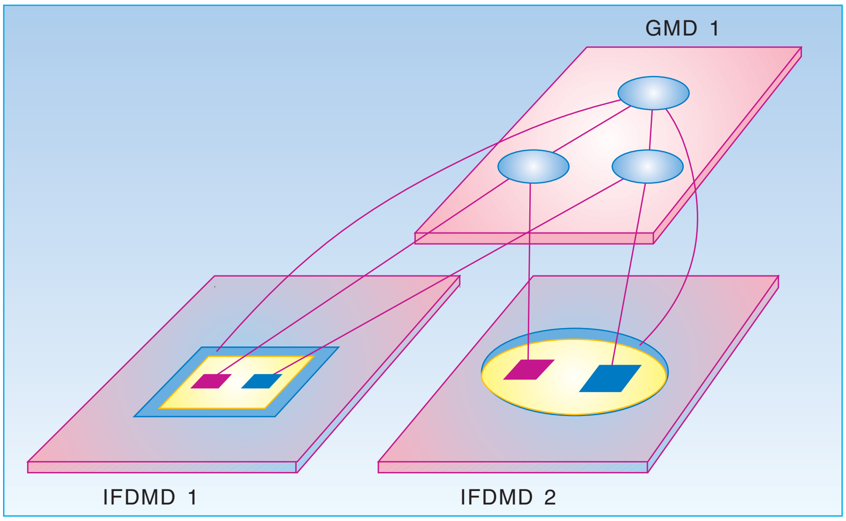


Figure 13 Mapping of one GMD to two different IFDMDs

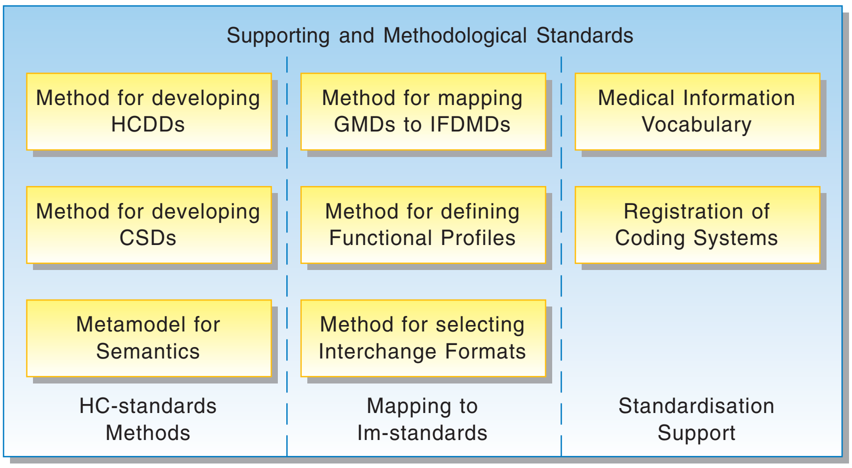


Figure 14 Supporting and Methodological Standards

7 Supporting and methodological standards

A number of supporting standards are required to establish the HC-standards and Im-standards. Figure 14 shows those which are currently identified in CEN/TC 251.

The standards are separated in three categories: Those supporting the establishment of HC-standards, those supporting mapping from HC-standards to Im-standards and a set of general supporting standards. The methods related to mapping have already been presented in chapter 6.1 and 6.2.

7.1 Methods for defining HC-standards

Development of HC-standards requires a standard methodology in order to ensure consistency across different HC domains. When developing health care messages, many methods and techniques may be used to describe the HC-domain and related scenarios. PT 004 (Investigation of syntaxes for existing interchange formats to be used in health care) has developed an initial method for developing HCDDs and CSDs. Object-oriented analysis is used as the description technique for DIMs and GMDs (Coad, 1991). The work is later taken over by PT 008 (Messages for exchange of laboratory information) which is trying the method in its attempt to establish messages for clinical laboratory medicine.

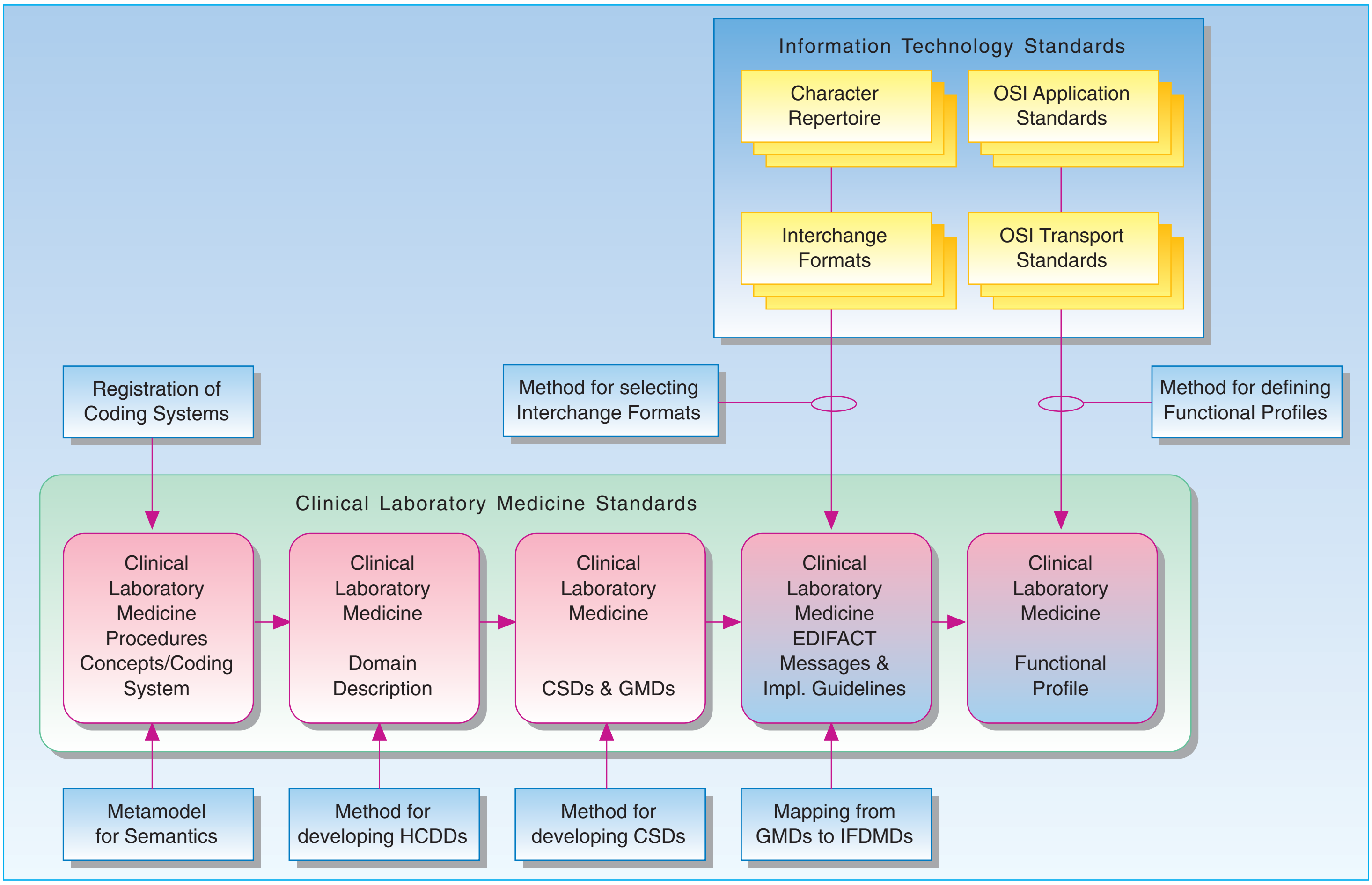


Figure 15 Standard for messaging in clinical laboratory medicine

Today's structure of coding systems is related to the manual use of the coding systems. Computers allow for a more flexible way of relating concepts and codes. A model for representation of semantics in medicine, developed by PT 003, describes a method for defining concepts and related classification and coding systems (CEN, 1993b).

7.2 Standardisation support

A Medical Informatics Vocabulary is invaluable to the co-ordination of work among medical informaticians. PT 001 is undertaking the job to produce such a vocabulary.

There is a need to uniquely identify existing and future coding systems to be used in health care. PT 005 is producing a standard which specifies procedures for the registration of coding systems.

8 Results in clinical laboratory medicine

Clinical Laboratory Medicine (CLM) has been the HC domain which has received most attention in CEN/TC 251. It was selected as a paradigm by WG 3 for the development of electronic messages in health care. Many of the supporting and methodological standards were identified and developed in order to establish CLM messages. Figure 15 shows the standards which are related to this domain.

The standards in rounded boxes are specifically related to the CLM domain. These are:

- Laboratory procedures, concepts and coding System (CEN, 1993c)
- The domain description of CLM
- Communication scenario descriptions related to CLM
- EDIFACT messages (MEDREQ and MEDRPT) and implementation guidelines for CLM
- Functional Profiles of OSI to support communication

The nerve project for establishing these sets of standards is the PT 008 (Messages for exchange of laboratory information). PT 008 has based its work on output from PT 004 and PT 005. Figure 15 shows the supporting, methodological and IT-standards related to CLM specific standards. Table 2 shows the relationship to the CEN/TC 251 Work Programme and the work items.

At present, the following objectives have been reached:

- The result of the PT 005 (Procedure for Registration of Coding Systems Related to Health Care) has been submitted for formal vote as a European pre-standard (ENV) (CEN, 1993d).
- PT 004 (Investigation of Syntaxes for Existing Interchange Formats to be used in Health Care) has produced a technical report (TR) which has been approved. The TR includes a draft method for developing HCDDs and CSDs, a method for selecting IFs and an evaluation of 5 IFs (CEN, 1993c).
- EWOS EG MED has produced a method for defining Functional Profiles (EWOS, 1993).
- PT 008 has finished a first working document for messages in CLM which has been distributed for comments. A final standard in CLM is expected in the spring of 1994.

9 Conclusions

This article has introduced a method for structuring health care telematics standards along two dimensions, health care and technology. Four categories of standards were identified: HC-standards, IT-standards, Im-standards, and SM-standards. The work in CEN/TC 251 was presented using this method.

At first glance, the defined method may look complex because of the number of categories and individual standards which are introduced. However, the health care telematics field is rather complex, taking the two dimensions into consideration. Separation of the HC-standards from the more technical standards should reduce the barrier for health care professionals to participate in the standard making and quality control process. It is a challenge for CEN/TC 251 to present its result in such a way that it could easily be understood by the user community.

Many small and well scoped standards should be better than few, large and unclear standards. This requires co-ordination and a common framework to get consistency among the standards.

New medical informatics terms have been introduced in the CEN/TC 251 work. However, the concepts behind these terms have often been explored by other standardisation groups. The concept of an object-oriented information model was first introduced by MEDIX. The concept of scenarios is under development in JTC 1/WG 3 Open-Edi. The information content of the CLM domain description is collected from standards such as ASTM 1238, EUCLIDES, EMEDI, HL7 and various national laboratory specifications. The reuse of other groups concepts is important for getting broad consensus on the resulting standards.

Standards are of no use if they are not implemented. The establishment of parallel implementation projects to the standardisation process where standards can be tested, is important. A special organi-

Table 2 Overview of clinical laboratory medicine and supporting standards

No	Standard Description	Category of Standard	Work Item
1	Metamodel for Semantics	SM-standard	WI 2.12 MOSE (PT 003) Model for Representation of Semantics
2	Registration of coding systems	SM-standard	WI 3.8 RCS (PT 005) Procedure for Registration of Coding systems of medical procedures
3	Clinical Laboratory Medicine - Laboratory Procedures - Concepts and Coding system	HC-standard Concept/Coding System	WI 2.4 PROCTERM (PT 002) Terminology and coding systems of medical procedures
4	Method for developing Domain Descrsiptions	SM-standard	WI 3.7a (PT 004) Investigation of Syntaxes for existing Interchange Formats to be used in Health Care ----- WI 3.13 METHODOL Methodology for the development of Healt Care Messages
5	Clinical Laboratory Medicine - Domain Description	HC-standard Helth Care Domain Description	WI 3.5 LABMES (PT 008) Messages for exchange of Laboratory Information
6	Method for developing Communication Scenario Descriptions	SM-standard	WI 3.7a (PT 004) Investigation of Syntaxes for existing Interchange Formats to be used in Health Care ----- WI 3.13 METHODOL Methodology for the development of Healt Care Messages
7	Clinical Laboratory Medicine - CSDs & GMDs	HC-standard Communication Scenario Description	WI 3.5 LABMES (PT 008) Messages for exchange of Laboratory Information
8	Method for Selecting Interchange Formats	SM-standard	WI 3.7a (PT 004) Investigation of Syntaxes for existing Interchange Formats to be used in Health Care
9	Mapping rules for GMDs to IFDMDs, EDIFACT	SM-standard	WI 3.5 LABMES (PT 008) Messages for exchange of Laboratory Information
10	Clinical Laboratory Medicine - EDIFACT Messages and Implementation Guidelines	Im-standard IFDMD and Implementation Guidelines	WI 3.5 LABMES (PT 008) Messages for exchange of Laboratory Information and WEEB MD 9 Health Care
11	Method for defining Functional Profiles	SM-standard	EWOS EG MED WI 3.1 OSI-APPS OSI Application Profiles for Health Care WI 3.2 OSI-TRANS OSI Transport Profiles for Health Care
12	Clinical Laboratory Medicine - Functional Profile	Im-standard	EWOS EG MED WI 3.1 OSI-APPS OSI Application Profiles for Health Care WI 3.2 OSI-TRANS OSI Transport Profiles for Health Care

sation, ACOSTA, has been set up on the European level to co-ordinate R & D, standardisation and industrial promotion in the field of health care telematics. Similar co-ordination is required on national levels involving the national health authorities and user-groups in implementation projects.

CEN/TC 251 has made rapid progress since its establishment in 1990. Today, about 700 individual experts are active in CEN/TC 251 including national mirror groups. Two standards are in the process of formal voting. At the beginning of 1994 a complete set of standards in the CLM domain are expected. Standards in the other domains vary from 1994 to 1996 depending on whether paid project teams are used to carry out the work.

CEN/TC 251 represents a new construction in the formal standardisation world. It serves a specific user-community (health care). Standardised user requirements are given to the technical standardisation bodies to influence the development of IT-standards. In the future it is likely that other industry sectors will create similar standardisation mechanisms as CEN/TC 251. Even though the resulting standards of CEN/TC 251 are not transferrable to domains outside the health care sector, the methods developed are fairly general and could easily be exported. CEN/TC 251 may then be a paradigm for a new category of user-oriented IT standardisation.

References

CEN. 1993a. *Directory of the European standardisation requirements for health care informatics and programme for the development of standards.* Gent, CEN/TC 251. (Draft version 1.7.)

CEN. 1993b. *Model for representation of semantics in medicine.* Gent, CEN/ TC 251. (Document 1993/N 5.)

CEN. 1993c. *Investigation of Syntaxes for Existing Interchange Formats to be used in Health care.* Gent, CEN/TC 251. (Document N93-87.)

CEN. 1993d. *Health care informatics interchange, Registration of coding schemes.* Gent, CEN/TC 251. (Document N93-44.)

CEN, 1993e. *Structure for classification and coding of Laboratory Procedures,* CEN/TC 251. (Document N93-3.)

Coad, P and Yourdon, E. 1991 *Object-Oriented Analysis.* 2nd edition. Englewood Cliffs, New Jersey, Prentice Hall. ISBN-0-13-629981-4.

De Moor, G. 1993. Standardisation in Health Care Informatics and Telematics in Europe: CEN TC 251 Activities. In : De Moor et al. *Progress in Standardisation in Health Care Informatics.* Amsterdam, ISO Press, 52-59. ISBN 90-5199-114-2.

EWOS. 1993. *A Method for defining functional profiles for health care.* Brussels, EWOS EG MED. (Document EWOS/EG MED/92/26.)

From, S, Stenvold, L. A, Danielsen, T. 1993. Telemedicine services integrated into a health care network. *Telektronikk* 1, 12-22.

ISDN: New possibilities for telemedicine

BY SIGMUND AKSELSEN, ARNE KETIL EIDSVIK AND TRINE FOLKOW

621.39:61
621.39.05

1 Introduction

The Norwegian Telecom Research (NTR) telemedicine project has led to several telemedicine applications (Nymo and Engum 1990; Gammon 1991). Today, the applications are based on a variety of networks, ranging from the ordinary telephone network to specialised data and video communication networks. The cost, ease of use and availability of network solutions and terminal equipment have been identified as major factors that have to be addressed in order to make telemedicine a realistic and affordable tool for the region's health institutions. Thus, the technical solutions pose limitations due to their

- cost
- lack of flexibility
  - design for special purpose use
  - special subscription requirements
- availability

Integrated Services Digital Network (ISDN) will provide one single network for a wide range of applications. An internationally standardised and widespread ISDN will also stimulate the production of a wide range of relatively cheap terminal equipment. In this paper we discuss the use of ISDN as an infrastructure for telemedicine applications.

2 Development of telecommunication networks

ISDN is an acronym for Integrated Services Digital Network, and is a further development of the telecommunication network. Telephony has until now been the dominating telecommunication service, and will probably be that also in the future. The data services, however, are expected to get increased importance. The dominance of telephony is reflected in ISDN, which originates from analogue telephone networks. As the network is digitised the possibilities for transmitting data is improved. Digitalisation and integration of services are the benefits of ISDN. Speech and other originally analogue information is converted into digital form already in the terminal equipment, is transmitted in digital form, and is converted back into analogue form in the receiver's terminal equipment. Integration of services implies that different services can be offered through one single subscription and "socket in the wall".

2.1 Telecommunication services offered by an ISDN

According to CCITT (1989), telecommunication services is grouped into bearer services, teleservices and supplementary services. A bearer service is used to transmit information through the network. A teleservice is the service a user meets in front of his terminal. A supplementary service can not be used on its own, but is always accompanied by a teleservice or a bearer service. Figure 1 illustrates the relationship between the services and gives an example of the use of the supplementary service Advise of Charge (AOC).

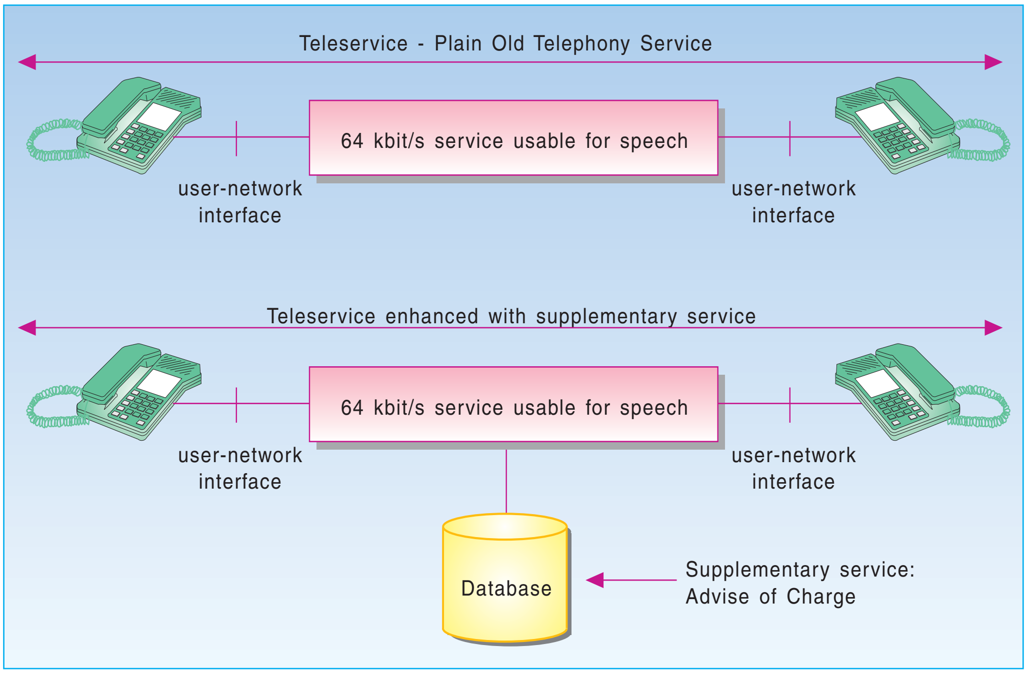


Figure 1 An example of telecommunication services (Akselsen et al 1991)

The following bearer services will, according to (CEPT 1992), be implemented in the Norwegian ISDN:

- 64 kbit/s unrestricted bearer service
- 64 kbit/s service usable for 3.1 kHz audio information transfer
- 64 kbit/s service usable for speech
- 2 * 64 kbit/s unrestricted bearer service
- Packet mode bearer service

Teleservices which will be offered include:

- Telephony 3.1 kHz
- Group 4 Telefax
- Telephony 7 kHz
- Audio Teleconferencing
- Videotex (alpha geometric mode)
- Videotex photographic mode
- Teleaction
- Videophone
- Messaging services

Supplementary services which will be offered include:

- Advise of charge services (AOC)
  - AOC, charging information at call set-up time
  - AOC, charging information during the call
  - AOC, charging information at the end of the call
- Number identification services
  - Calling line identification presentation (CLIP)
  - Calling line identification restriction (CLIR)
  - Connected line identification presentation (COLP)
  - Connected line identification restriction (COLR)
- Closed user group (CUG)

- Call waiting (CW)
- Completion of call to busy subscribers (CCBS)
- Conference services
    - Conference call, add-on (CONF)
    - Meet me conference (MMC)
- Direct dialling in (DDI)
- Diversion services
    - Call forwarding unconditional (CFU)
    - Call forwarding busy (CFB)
    - Call forwarding no reply (CFNR)
    - Call deflection
- Freephone (FPH)
- Malicious call identification (MCI)
- Multiple subscriber number (MSN)
- Subaddressing (SUB)
- Terminal portability (TP)
- Three party service (3PTY)
- User-user signalling (UUS)

ISDN provides various telecommunication services through one common interface to the subscriber line. Previously, access to these services were possible via special add-on equipment and over separate networks.

Due to the multiplicity of services it is unreasonable to restrict the customers to use only one channel at a time. For this reason a basic rate ISDN access, offering two 64 kbit/s channels, is standardised, thus allowing the use of more than one terminal at a time. These channels are called B channels. A 16 kbit/s signalling channel is also offered. This channel is called a D channel and may be used for user-to-user information as well as signalling. Packet mode communication is facilitated on both B and D channels, whereas circuit mode communication is facilitated on the B channels. It is possible to connect 8 different terminals in parallel across the incoming subscriber line. A primary rate ISDN access offers 30 B channels (23 in the US) and 1 D (64 kbit/s) channel. In addition there exists terminal equipment that can combine several B channels to increase capacity (n * 64 kbit/s).

The telecommunication services offered by ISDN must be accessed by digital terminal equipment. This equipment includes:

- Telephones
    - alphanumeric keyboard and display for simple text messages
    - 3.1 kHz telephony with enhanced performance (quality of digital transmission, quicker call set-up, new supplementary services)
    - 7 kHz telephony offers better voice quality and is appropriate for broadcast-quality calls and audio teleconferencing
- Videophones
    - simultaneous transmission of speech and moving pictures
- Group 4 Telefax
    - transmission of fax messages with enhanced quality (400 * 400 dots per inch) and speed (one A4 page in less than 10 s)
    - colour transmission under development
- PCs with cards for voice, data, and video communication
    - file transfer
    - messaging
    - desktop conferencing
    - videophony

In addition it is possible to connect existing (pre-ISDN) terminals to ISDN through terminal adapters.

2.2 ISDN Features

Digital transmission and use of a common network for different kinds of services offer the users new possibilities:

- Standardisation of services, user procedures, terminals and access line connections/interfaces
- Integration of services including voice, data and picture communication
- Availability of a large number of services through one "socket in the wall"
- Flexibility when it comes to use of channels, choice of bearer services, choice of telecommunication service, and connection of terminal equipment
- Reasonable costs. Taxation is based on volume of transmitted information when using packet mode transmission or on connection time when using circuit mode
- High transmission rates (e.g. faster and better facsimile and other text services)
- Less error-rate on transmission
- Secure transmission of information (e.g. functions such as closed user group, prevention against tap of information and guarantee for transmission of identity by the network).

ISDN has been standardised internationally by CCITT (Kano et al 1991). Today, ISDNs are implemented and accessible on a limited scale. Within 5 years, however, ISDN is expected to become widely available in most countries in Europe, North America and East-ern Asia.[1)]

For the health services the ISDN infrastructure offers the opportunity to solve most telecommunication needs (Akselsen et al 1991). Reasonable costs and one standardised network may make applications available for institutions on a large scale.

2.3 Further development

The development of a digital telecommunication network leads to an infrastructure that makes it possible to introduce intelligent networks (IN). IN consists of methods and tools for defining, developing, testing, and implementing customised services. Centralised databases and service control are essential parts of IN. Typically, information from the databases will be used for controlling resources in the network. When an incoming call is detected in the exchange, further actions are suspended until relevant instructions are retrieved from a database. This facilitates an independent control of services activated from several different networks. IN is discussed in more detail in (Løken 1992).

Some examples of the services that will be available through IN are:

- universal personal telecommunication service
- queue service

1) *Notice for instance the launch of Euro/ISDN in December 1993, which will provide a common ISDN between a number of European countries, including among others Belgium, Denmark, Germany, France, Ireland, Italy, Netherlands, Norway, Portugal, Spain, Sweden, and the United Kingdom.*

Table 1 Telemedicine applications and possible ISDN network solutions

Application	ISDN solution
Telediagnosis	
Radiology	BRA
Pathology	PRA
Endoscopy	PRA
Gastroscopy	PRA
Dermatology	PRA
Microbiology	PRA
Auscultation	BRA
Electrocardiology	BRA/GSM
Electroencephalogy	BRA/GSM
Groupware for making diagnoses	BRA/PRA
Distance learning	
Aphasic patient rehabilitation	BRA
Decentralised nursing education	BRA/PRA
Medical information	
Medical information service	BRA
Laboratory communication	BRA
Patient record transmission	BRA
Safety alarms	BRA
Patient monitoring	BRA
Administrative information	
On-line booking	BRA
Computer-supported telephony	BRA/PRA
Videophones and video conferences	BRA/PRA
ISDN networks for data communication	BRA/PRA
Integration of fixed and mobile networks	BRA and GSM

GSM - Global System for Mobile communication
BRA - Basic Rate Access
PRA - Primary Rate Access

- credit card calling
- televoting
- green number service (freephone)
- premium rate service
- alternate billing service
- emergency response service
- private virtual network
- area wide centrex.

The mobile phone market is developing rapidly. The capacity of analogue transmission is limited. By using digital transmission, it will be possible to increase the capacity considerably (Valentino 1992). The European digital cellular telephony, known as Global System for Mobile communications (GSM), is defined to be a land mobile system for voice and data services using digital cellular radio technology for the connection to the public switched telephone network, to ISDN, and to the public data network. All bearer and teleservices offered today will in principle (with limited capacity) be offered by GSM. Furthermore, GSM will provide mobile equipment and pan-European infrastructure at reduced costs.

3 Applications

ISDN will increase the availability of some applications, improve the quality of some and facilitate some completely new ones. The health services promise a large market potential for cost-effective telecommunication services. ISDN will, with a variety of bearer services, be cost-effective and within range for both large and small institutions. Thus, there will be a large market for applications designed particularly for the health services.

We will in the following summarise applications generated from the NTR project "Telemedicine in North Norway". ISDN applications may be categorised according to several criteria. One way is to group applications into the following main categories reflecting the type of information transmitted (Guenin et al 1991; Zulke and Chopard 1990): image, documentation, data and multimedia. For our purpose we have chosen to group applications according to their use in the health service (Akselsen et al 1993): telediagnosis, distance learning, medical information and administrative health information.

Table 1 gives an overview of some telemedicine applications in Norway. ISDN is considered as a network solution for all of these. Some of the applications are already implemented with other network solutions. Below, some examples of the applications will be described in further detail.

3.1 Telediagnosis

3.1.1 Telepathology

Telepathology is the performance of pathological services at a distance using telecommunications (Nordrum et al 1991). Pathologists examine samples of tissue using microscope. For diagnosing some kinds of cancer, the pathology service must be performed while the patient is still anaesthetised after the sample of tissue is removed, in case further surgery is needed instantly.

Telecommunications are at present based on a 2 Mbit/s connection between the sites. An ISDN solution will be more cost-effective. Packet mode communication can be used for remote control of the microscope and circuit mode for images and sound.

3.1.2 Teleradiology

Teleradiology is the practice of radiology at a distance using telecommunications. This involves transmission of, for instance, X-ray, computer tomography (CT) and ultrasonic images. At present the data communication is implemented with a 64 kbit/s permanent connection and Vitalink bridges connecting LANs at the two hospitals. Equipment and project are further described in (Sund et al 1991a; Sund et al 1991b).

The first step towards an ISDN implementation of teleradiology will be to replace the Ethernet bridges with IP (Internet Protocol) routers for ISDN, and the 64 kbit/s permanent connection with a basic rate ISDN subscription. X-ray images comprise large amounts of information and all images for one day are usually transmitted in one batch. Large volumes of data have to be transmitted, and a circuit switched connection will be most suitable. Hence, users will pay only for connection time. Loss of information is an important concern in compression of medical images, hence the relatively large capacity of ISDN is needed. ISDN communication over IP routers will in addition provide a transparent LAN to LAN connection between the two hospitals, transparent in the sense that users at one of the LANs can access services/resources at the other one as if they were on the same (local) network.

3.1.3 Tele-endoscopy

An endoscope is a device for guiding a source of light inside the human body and transposing an image of the examined organ on a monitor. Tele-endoscopy refers to the transmission of these images over the telecommunications network to a specialist at the other end. A practitioner is, however, still needed to bring the endoscope inside the patient and adjust the positioning to get the correct images (Hartviksen and Pedersen 1992).

Tele-endoscopy is used in North Norway for ear, nose and throat (otorhinolaryngology) consultations. Telecommunications are today implemented with a 2 Mbit/s switched connection. Video conferencing equipment is used for presenting images and transmitting speech between the specialist and the practitioner in charge of the patient. An ISDN implementation may replace the 2 Mbit/s connection. In fact, the initial trials using the Norwegian Tandberg Vision videophone (Møllerbråten 1991) over 2 ISDN B-channels, i.e. 128 kbit/s, has already been done. An ISDN solution where several B-channels are combined, e.g. 6 channels which total 384 kbit/s may be necessary if the quality of 128 kbit/s is too poor. An alternative/complementary system for making remote diagnoses is under development. This system can digitise, store and transfer medical stills (pictures) from arbitrary video sources. Among other medical areas, gastroscopy is considered as a suitable area for making remote diagnoses using similar tools.

3.1.4 Groupware for making diagnoses

Multipurpose groupware can be used by medical practitioners for co-operation and consultation when making diagnoses. ICL's Desktop Conferencing application, designed particularly for ISDN, is an example. Eight users may, in addition to voice communication, share screen images and applications. An extension of the system to include videophony will make it possible to utilise the groupware facilities further. The specialists will be able to look at pictures, hear examples of sounds and read medical records to get an overview of the patient in question. Participants of the co-operating group may point at, and modify, the screen images as the discussion goes on. One participant is assigned to be the chairman. He is supported with computer based means for directing the meeting, e.g. dedicate the cursor control to each participant in turn.

3.2 Distance learning

Distance learning and education has evolved from the use of paper, pencil and ordinary mail to data communication and videoconferencing. ISDN has the facilities to combine synchronous (videoconferencing, groupware, etc.) and asynchronous (electronic mail, electronic conferencing) communication. Both institutions and private households will benefit from distance learning over ISDN due to low costs. Distance learning may thus be used for educating patients and clients as well as health care personnel. The following project from the NTR telemedicine portfolio illustrates how ISDN may be used for patient rehabilitation and learning.

3.2.1 Aphasic patient rehabilitation

Aphasia is loss or reduction of the ability to use and understand language, often caused by stroke (apoplexy) or injuries from an accident. People suffering from aphasia have to learn their own language all over again. In our project the speech therapists and the clients were supplied with personal computers configured for synchronous communication. Graphics and text containing exercises were transmitted from the speech therapists' computers to the clients' computers over ordinary telephone lines. The clients worked with the exercises in their own spare time or assisted by on-line guidance from the speech therapist. When online, therapist and client had access to the same screen image on the computers, facilitating pointing and editing. In addition voice communication was offered over the same connection (Eidsvik 1990; Holand 1991).

This application is well suited for ISDN, as the combination of data and voice communication is essential. At present the system uses ordinary telephone lines with modems facilitating both data and voice communication on the same line, resulting in poor quality and low transfer rates. ISDN will give much better audio quality and much faster data transfer. Supplementary ISDN services such as Calling Line Identification Presentation (CLIP) may simplify the user interface for the patients, with for instance automatic start of the software upon a call from the speech therapist.

3.3 Medical Information

3.3.1 Medical information service (MEDIS)

Medical research and development generates enormous amounts of information. The information may be retrieved from both traditional literature and various electronic sources. User interfaces are different for most sources of information. Medical practitioners will experience an increasing problem of keeping up with new results and methods. The project MEDIS is aimed to give users a uniform user interface to several sources of information including electronic mail, electronic conference systems and on-line databases (Danielsen et al 1990; Wasson and Akselsen 1992).

Higher rates of data transfer through ISDN will facilitate a better user interface compared to the possibilities of ordinary telephone lines. The integration of voice and data communication in ISDN, and two or more parallel communication lines for each subscriber, makes it possible to combine database searching with add-on information from human experts.

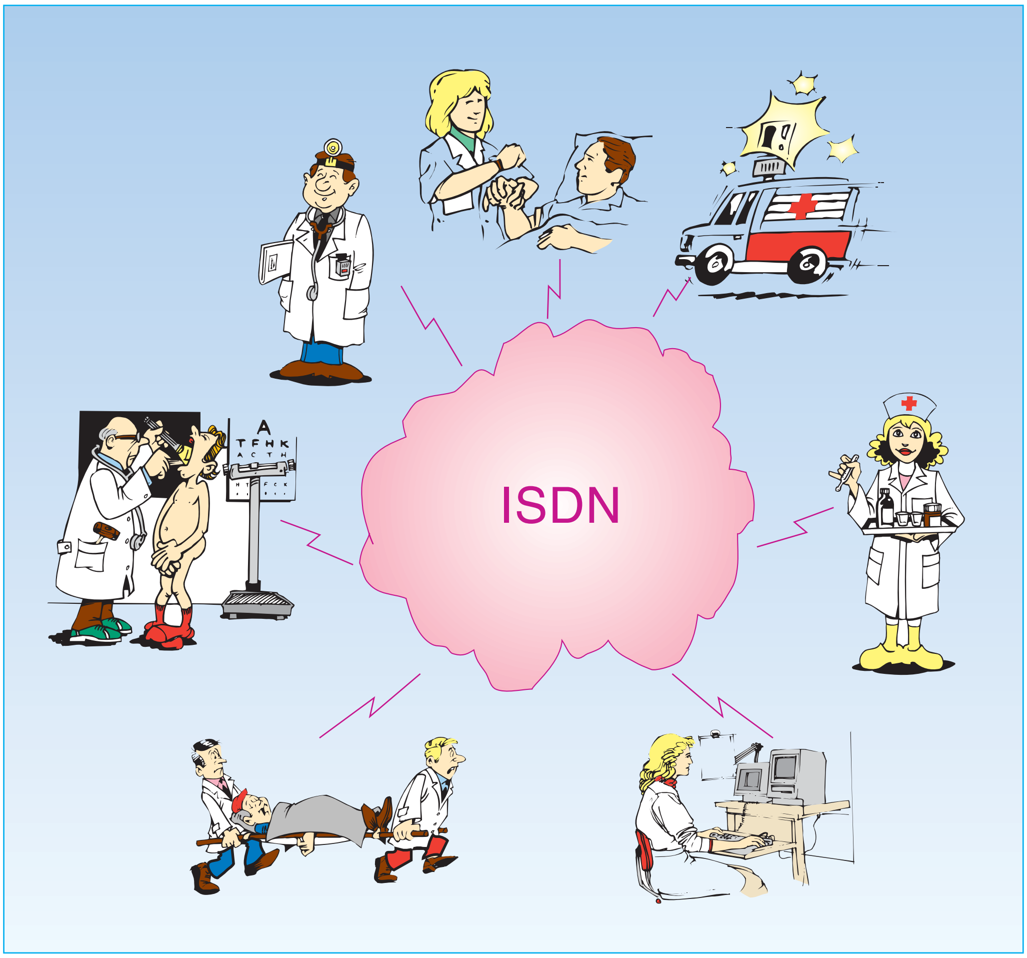


Figure 2 Telecommunication for exchange for medical information (Akselsen et al 1993)

3.3.2 Laboratory communication

Ordering and transmitting laboratory results between clinical laboratories and customers is usually done through the postal service. Several laboratories offer electronic result services based on various models for data communication. In our project solutions have been offered, ranging from telefax generated by the clinical test computer, to electronic mail based on TelemaX.400[2]. Investigations have been conducted on exchange of medical information based on Electronic Data Interchange (EDI), see e.g. (Brox and Antonsen 1990).

Today, the data communication is implemented with modems, ordinary telephone lines and packet switched WANs. With ISDN, transfer rates will be higher and the interface between network termination and user terminal simpler.

3.3.3 Patient record transmission

Multimedia patient records may be transmitted over large distances, between small and large institutions, with ISDN. Such records may in addition to text include pictures and examples of sounds (from breathing, heart beating, etc.). Sounds and pictures may be used to show the development of the disease over time. This kind of information combined with speech transmission will give practitioners a valuable tool when discussing medical cases.

For years to come, many patient records will still be available on paper only. Group 4 Telefax offering high resolution printouts will be a suitable medium for transmitting patient information presented by text and pictures.

2) *TelemaX.400 is an X.400 electronic mail service offered by the Norwegian Telecom.*

3.3.4 Safety alarms

Safety alarm users can rely on help and comfort, night and day, wherever they are in their home. If help is needed, the emergency unit is alarmed by a push on the alarm button of a radio transmitter carried by the patient. The transmission range of a typical transmitter is approximately 60 metres, depending on building constructions. The system therefore requires that the patient is within the range of the receiver, and that he does not lose consciousness. These requirements evidently limit the use and value of the safety alarm.

The limitations of existing systems could be eliminated by combined use of a patient monitoring system providing real time diagnosis and tracking of dynamic physiological events (e.g. ECG, haemodynamics, EEG), a radio transmitter and ISDN (alternatively the mobile counterpart GSM). The patient can either alarm the emergency unit manually, or the alarm can be triggered by the monitoring system. In either case the CLIP can be used by the hospital's application program to find the patient's file with information about where he lives, what kind of disease he suffers from, etc. If the patient is not at home and unable to explain where he is, a distress signal can be used to locate him.

This system configuration gives the patient freedom to move. It may enable the hospital to give help sooner, and thereby increase the possibility to save lives, e.g. in cases such as sudden infant death and heart, asthma or epilepsy attacks.

3.4 Administrative information

3.4.1 On-line booking

Today, practitioners book hospital treatment for their patients by writing applications to the hospital department in question. Electronic equipment and modern telecommunications make it possible to facilitate on-line booking. The practitioners can book treatment at hospital departments with the patient present (just as the customer at a travel agency). The patient may leave his physician knowing the exact date and time for his appointment. Necessary information can be transmitted between the practitioners electronic records and the booking system.

Using ISDN as the communications infrastructure, where the taxation rates depend on mode of transmission, will be cost-effective when searching the booking system (using packet mode) and transmitting data (using circuit mode). In addition, telephony may be used simultaneously for consultations with administrative personnel in cases of doubt.

3.4.2 Computer-supported telephony

Supplementary services can be used to enhance the basic services in ISDN. An ISDN will offer a number of supplementary services, e.g. call waiting, direct dialling in, closed user group. Among these services, CLIP provides an interesting feature to support the telephony services of health care institutions. A CLIP can be used by the called subscriber's application program to take automatic actions (Iffland et al 1989). This feature can be utilised when: individuals phone health institutions; health institutions phone each other, and; health institutions phone public or governmental institutions.

The receptionist at a primary health care clinic can, for instance, use a program that aids the identification of a person/family from the CLIP, and retrieves the medical record in question. If a practitioner is to be consulted, the medical record can be transmitted to the practitioner's terminal together with the telephone call. In a similar manner, institutions can be identified by CLIP to automatically retrieve relevant information. The communicating institutions can for example include: a laboratory and a primary care clinic; a pharmacy and a primary care clinic, or; a Social Security office and a primary care clinic.

3.4.3 Videophones and video conferences

Videophony and video conferencing over ISDN will be useful tools for communication within the health services (see e.g. Tetzchner and Holand 1991; Holand, Tetzchner and Steindal 1991). In many of the remote diagnosis and distance learning applications, videophony or video conferencing is essential. It may also be a useful tool when transmitting medical information in the form of video sequences. In administration expenses may be reduced by arranging some meetings through video conferences instead of travelling. ISDN will facilitate videophony and video conferencing implemented through a flexible telecommunication network that can be used for a number of purposes.

3.4.4 ISDN networks for data communication

Enhanced features for data communication will be a particularly important service of an ISDN portfolio. In the health services, use of ISDN for the exchange of administrative information has a great potential. Making an appointment for an X-ray examination (radioscopy) may suit as an example. The primary health care clinic's system for medical records can establish a connection to the local hospital's booking system while the patient is still at the practitioner's office. Then the patient, the practitioner and the clerical officer at the hospital co-operate (via speech on a B-channel) on deciding a suitable time and make the appointment. The necessary information on the patient can then be transmitted. After a completed examination, the X-rays and the radiologist's comments (diagnosis) can be transmitted back to the primary health clinic system for medical records.

In addition, there exist some possibilities for providing general data communication over ISDN. The network solutions include:

- Basic rate access as a local "mini" computer network, e.g. with PCs sharing a printer on the same passive S-bus. This solution might be interesting for internal communication in primary health clinics.
- ISDN as a WAN PC-network where PCs, file servers and printers are interconnected over a distance. This solution might be interesting for communication between departments within a hospital, between primary health care clinics and centralised institutions.
- Interconnection of LANs through gateways (e.g. IP-routers) over ISDN. The idea of having all parts of the (public) health services on the same LAN is interesting. Given proper authorisation, everybody can access every resource in every network as if they were on the same LAN, it be printers, plotters, databases, files and process servers.

3.4.5 Integration of fixed and mobile networks

The future integration of fixed (e.g. ISDN) and mobile (e.g. GSM) networks makes new applications possible for telemedicine far beyond use of today's mobile telephony. A scenario might be the practitioner on a home visit. With the help of his mobile multiservices telecommunication terminal he retrieves (and updates) information from the patient's medical record at the local hospital, consults a specialist with accompanying pictures of the patient, and finally places a description for the patient's medication in the local pharmacy.

4 Discussion

Above we have given some examples of the new possibilities ISDN bring. For the users ISDN will offer:

- integration of services
- flexibility with respect to use of channels, bearer services, teleservices and terminal equipment
- secure transmission of information
- high transmission rate
- stable and high quality connections
- standardised services, access points, procedures and terminals
- reasonable costs.

The hospitals will in particular benefit from advanced and functional communication services within and between institutions. Small institutions and the primary health care will get an increased repertoire of telecommunication services and may connect up to eight terminals to one single basic rate access. ISDN will provide an infrastructure for communication, both within and across the hierarchical levels in the health service (see figure 3). In most cases one or a limited number of basic rate accesses will satisfy the communication needs for units in the primary health service. Between hospitals and larger institutions more bandwidth will be necessary for certain applications. Thus, primary rate access will be appropriate.

Standardisation is important for a widespread international ISDN were communications can take place across national borders. This makes it possible to offer telemedicine services on an international basis. As a matter of fact requests for such services have already been received by the University Hospital of Tromsø from the northern parts of Sweden (and even Kuwait). It will also be possible to offer telemedicine services from the industrialised world to developing countries. Furthermore, ISDN offers a unique opportunity to build interna-

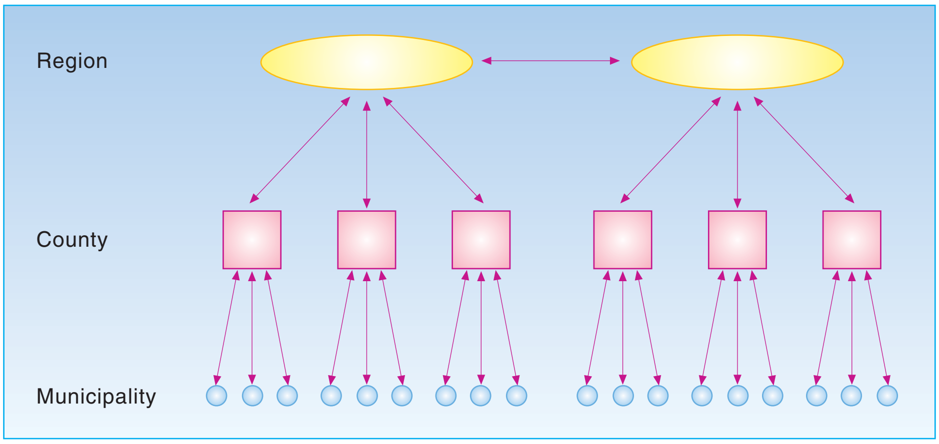


Figure 3 Communication structures in the Norwegian health service hierarchy

tional networks for professional discussions and transfer of competence.

However, there are some factors that may prohibit the promises that ISDN brings to telemedicine. The current costs of terminal equipment and lack of compatibility limit the number of ISDN users. As with any new technology, ISDN also is dependent on a critical mass of users. If this mass is not reached within a reasonably near future, ISDN may become just a parenthesis in the history of telecommunications.

In the health service compression of medical data is controversial. Thus, for some telemedicine applications the capacity of ISDN communication may not be sufficient.

5 Summary and conclusions

In a five year period ISDN technology and services will be available in most industrialised countries. ISDN will offer new services and higher transmission rates at lower prices. ISDN also will provide greater flexibility with regard to existing communication solutions. Until now, the utilisation of ISDN has been tied mainly to already existing services and applications. However, developments are still at an early stage when it comes to utilising calling-line identification and other supplementary services. Furthermore, IN services are being developed and will become an important part of the future telecommunication services repertoire. These services will find numerous applications within the health service. One illustrative use of IN services is utilisation of the universal access number service to call a pool of specialists and get direct access to one available.

One of the arguments for the development of telemedicine services in Norway has been the principle of equal access to health services, independent of economic, social, religious, race and last but not least, geographical issues. A widespread ISDN will constitute an important part of the infrastructure necessary for fulfilment of this principle.

We consider the principle of equal access a most important argument in justifying developments in telemedicine. The future of telemedicine depends on applications that are technically and organisationally possible, economically feasible and politically acceptable. ISDN can fulfil the technical and economical conditions necessary for such applications.

We have reviewed some telemedicine applications and proposed a number of ideas for further development of such applications based on ISDN technology. We have argued that ISDN may increase the possibility for successful use of telemedicine. However, neither ISDN nor any other technology alone will automatically fulfil the goals of equality and decentralisation. A unified political strategy is a necessary precondition.

The availability and the relatively (cost-effective) low pricing of future ISDN services and terminals are the keys that make telemedicine realistic for the public, i.e. available to a majority of health care personnel and not just a curiosity for the "chosen few".

6 References

Akselsen, S, Eidsvik, A K, Folkow, T. 1991. *Health services and ISDN.* Kjeller, Norwegian Telecom Research. (TF-report R68/91.) (In Norwegian.)

Akselsen, S, Eidsvik, A K, Folkow, T. 1993. Telemedicine and ISDN. *IEEE Communications Magazine,* 31(1), 46-51.

Brox, E, Antonsen, T. 1990. *Exchange of medical information based on EDI.* Foundation for Applied Research at the University of Tromsø. ISSN 082-2011.

CCITT. 1989. *Integrated Services Digital Network (ISDN). General Structure and Service Capabilities.* Geneva. (CCITT Blue Book vol III - fascicle III.7.)

CEPT. 1992. *User Handbook ISDN.* Bern, CEPT (The European Conference of Postal and Telecommunications Administrations).

Danielsen, T et al. 1990. *RATATOSK - an adaptive user interface to computer-mediated communication.* Presented at "Message Handling Systems and Application Layer Communication Protocols", Zürich, 3-5 October. Also available as TF-lecture F16/90, Kjeller, Norwegian Telecom Research.

Eidsvik, A K. 1990. *Distance learning for people suffering from aphasia.* Presented at "IFIP Fifth World Conference on Computers in Education", Sydney, 9-13 July. Also available as TF-lecture F18/90, Kjeller, Norwegian Telecom Research.

Gammon, D. 1991. *Telemedicine in North Norway.* Kjeller, Norwegian Telecom Research. (TF-report R27/91.)

Guenin, J P et al. 1991. ISDN in France - 1987 to 1990: from the first commercial offering to national coverage of NUMERIS. *IEEE Communications Magazine,* 29(1), 30-35.

Hartviksen, G, Pedersen, S. 1992. Remote endoscopy of otorhinolaryngology patients. In: A Katzir (ed): *Optical fibres in medicine VII, SPIE Proceedings,* 1649.

Holand, U. 1991. Use of graphic communication in distance training of patients with aphasia. In: S v Tetzchner (ed): *Issues in telecommunication and disability.* COST 219.

Holand, U, Tetzchner, S v, Steindal, K. 1991. Two field trials with videophones in psychiatric and habilitative work. In: S v Tetzchner (ed): *Issues in telecommunication and disability.* COST 219.

Iffland, C, Norton, G, Waxman, M. 1989. ISDN applications: their identification and development. *IEEE Network Magazine*, 3(9), 6-11.

Kano, S, Kitami, K, Kawarasaki, M. 1991. ISDN standardisation. *Proceedings of the IEEE,* (Special issue on ISDN), 79, 118-124.

Løken, B (ed). 1992. Intelligent Networks. *Telektronikk,* 2.

Møllerbråten, B. 1991. Videotelephone development in Norway. In: T Kristiansen (ed): *A window to the future. The videotelephone experience in Norway.* Kjeller, Norwegian Telecom Research, 9-20.

Nordrum, I et al. 1991. Remote Frozen Section Service; A Telepathology Project in Northern Norway. *Human Pathology,* 22, 511-518.

Nymo, B, Engum, B. 1990. *Telemedicine to improve the quality, availability and effectiveness of the health service in rural regions.* Kjeller, Norwegian Telecom Research. (TF-lecture F10/90.)

Sund, T, Rinde, E, Størmer, J. 1991a. Full-scale replacement of a visiting radiologist service with teleradiology. In: H.U. Lemke et al (ed): *Computer assisted radiology, Proceedings of the international Symposiums CAR-91.* Berlin, Springer-Verlag.

Sund, T et al. 1991b. Techniques for multimedia Teleradiology. In: H.U. Lemke et al (ed): *Computer assisted radiology, Proceedings of the international Symposiums CAR-91.* Berlin. Springer-Verlag.

Tetzchner, S v, Holand, U. 1991. Supervision of habilitation and psychiatric work. In: T Kristiansen (ed): *A window to the future. The videotelephone experience in Norway.* Kjeller, Norwegian Telecom Research, 32-48.

Valentino, G. 1992. A survey on Cellular Mobile Communications Systems. In: *Proceedings of the twelfth international symposium on biotelemetry.* August 31 - September 5, Ancona, Italy.

Wasson, B, Akselsen, S. 1992. An overview of on-line assistance: from on-line documentation to intelligent help and training. *The knowledge engineering review,* 7(4), 289-322.

Zulke, M, Chopard, S. 1990. ISDN Applications in France. *Telecommunications,* 24, 45-48.

The challenge of computer-mediated communication in health care

BY THORE DANIELSEN

621.39:61
681.3.069

Could CMC turn out as a helpful tool for the exchange of information within health care? Could CMC support situations beyond the writing of a note or a short letter to a colleague or a co-operating institution? Could CMC become a tool for bridging the information islands? Islands which are a result of implementation of information systems designed to cover the needs of the individual. Bridging is anyway a repair process through the exchange of data between information systems. In this paper we shall first investigate electronic mail and other CMC applications. This will form a basis for a discussion on the potential and the pitfalls for CMC, computer-mediated communication, in health care. The discussion is partly based on surveys of communicational patterns.

Introducing CMC

The concept CMC will be used as a reference to both electronic mail systems (e-mail) and the more juicy groupware or CSCW systems. E-mail systems have been around long enough to have matured into commercial products in daily use by a number of people. Systems for CSCW (computer supported co-operative work), or groupware, are still in its infancy except for the more basic conference systems, i.e. vanilla flavoured groupware.

E-mail systems will be used as the basis for the first part of the paper when discussing usage, usability and effects of CMC. Groupware will be considered in the concluding discussion and visions for future use of CMC.

E-mail systems provide users with the tools necessary for creating, distributing and deleting messages or "electronic letters". As the name itself indicates, the functionality of such systems may be seen as a computer-based copy of traditional mail systems except for two major differences: speed and methods for handling message contents. These, and more subtle differences, all induce changes in interpersonal relations and inter-organisational behaviour.

Early CMC applications included methods for handling communication within groups of users, often based on the Delphi method. The first of such systems to be used was EMISARI way back in the early seventies (Hiltz 1977). Later followed systems like EIES, CONFER, COM etc. They differ from the distribution lists by two main aspects:

- they maintain open lists of users for each group which any other user may look into whether she is a member of the group or not, unless the group is closed
- they maintain a store of messages exchanged for a longer period of time.

Most attempts at introducing CMC have been based on the idea that this new medium could replace the discussion and conversation which earlier had to take place in face-to-face meetings, through the use of telephone or through the exchange of written letters. This idea of replacement, however wrong it might be, have turned out as the official policy. For, as McLuhan & Fiore points out, "In the name of "progress," our official culture is striving to force the new media to do the work of the old." (McLuhan & Fiore, 1967). We shall later return to this statement and show why this might actually slow down the introduction of CMC in health care environments.

CMC – spreading the message

Exact figures concerning the use of CMC are hard to find. An estimation done in the mid-eighties (Quarterman & Hoskins 1986) show above 200,000 users, i.e. mailboxes on the UUCP (unix-to-unix copy) mail network alone. These users of UUCP were served by some 7,000 hosts. The total number of mail hosts listed in the same survey is close to 30,000. One third of the hosts did however belong to the corporate network of Dec. IBM would supposedly also support a vast number of mailboxes, though their number of hosts was less than a quarter of what Dec had at that time. Most hosts in these networks did also support remote login and file transfer. In addition some 2,500 hosts gave more than 50,000 users access to USENET news, a distributed distribution list service (see below). And these numbers does not include any of the dial-up bulletin-boards systems. An estimate for today would show some million users on a world basis accessing USENET News through some hundred thousand hosts.

A forecast made in 1988 (Clausen 1991) indicates that the total number of public and private mailboxes throughout Northern America, Europe and Pacific Asia should exceed 50 million by 1993. The increase over the 5-year period from 1988 was then indicated as approximately 40 % per year. Electronic mail is in a unique position as it "is currently the only organisational computing tool in widespread use by many kinds of people in many kinds of organisations" (Sproull & Kiesler, 1991).

The success of any tool for communication relies on the sheer number of peers you may reach through the system. The 50 million mailboxes are by no means interconnected. Most mailboxes reside within private or corporate networks with lacking or no gateways to public systems. Currently close to 70 e-mail services which are listed in the monthly EEMA Briefing[1)] show a disappointing state-of-the-art for interconnections. Electronic mail is therefore no universal tool for interpersonal communication which is the one major deficiency when compared to telephony and ordinary mail. The power of the simplicity of new services did prove right when the telefax was introduced. Unlike electronic mail, telefax may be used with cheap and functional equipment on an existing network.

CMC have although reached a high level of acceptance and usage in academic environments. Most university and research institute employees in technology-oriented sciences and a few others have access to CMC-services. But again, the use of CMC seems to depend more on the tradition of the institution and the interpersonal network of the individual than of the ease of access to CMC. And matters does not necessarily get better when we move on from the interpersonal e-mail to the group-oriented conference tools.

CMC – efficiency gained or efficiency lost?

Speed is often described as the main advantage with electronic communication. Next, we often find the reduced need for paper and printed versions of the communication as reasons for reduced costs in interpersonal communication. The use of electronic group mail - groupware - may, as pointed out by Sproull and Kiesler (1991), lead to a reduction in the costs of co-ordinating groups of communicators. The asynchrony provided by the CMC system eases the task of scheduling and co-ordinating participants. But asynchrony does also determine communicational processes for

1)The monthly Newsletter of the European Electronic Messaging Association.

which CMC is not suited. As shown by Whittaker (Whittaker, 1991), there are "a number of underlying media factors which determine that asynchronous interaction can never be incremental". Hence, "asynchronous communication is inappropriate for various types of activity that rely on negotiation or shared meaning because these require incremental communication." (ibid). Asynchrony may also delay decision making. An evaluation of two face-to-face and computer conferencing techniques (Archer 1990) showed that "the elapsed time spent in arriving at a decision was generally much longer for CMC than for face-to-face". Such results should not be too unexpected from in-house experiments. But the same experiment did also show that the quality of the decision-making "did not depend on whether or not group interactions were computer-mediated" or not. (ibid.) Thus, CMC may, if it is applied to the "right" communicational processes, result in an increased efficiency due to speed and asynchrony.

The integration of CMC into the users' environments may however turn out to be crucial if the gained efficiency shall not be lost. So far, most CMC systems have been given minimal functionality and little or no integration with other information systems. The result is CMC systems suited for exchanging short messages and memoranda and less suited for exchanging documents given in users' electronic file. A few approaches have although been made at designing CMC agents capable of supporting links between electronic mail and the stored information, cfr. (Motiwalla & Nunamaker 1992). At the heart of these agents we find processing mechanisms known as knowledge-based (Motiwalla & Nunamaker 1992, Richardsen & Danielsen 1989). Another approach could be to look at the integration of all communication tools into one single environment. This is for instance found in the Ratatosk environment (Danielsen et al 1991, Finnset 1992, Hartvigsen 1992).

Efficiency is more than increased speed. New technology may prove to have effects beyond what is measured in traditional analyses. It may well be that "the most important effects of a new technology may be, not to let people do old things more efficiently, but rather to do new things that simply were not possible or feasible with the old technology" (Sproull and Kiesler 1991). These "second-level effects come about when a new technology connects people with new information and new people". (ibid.) Once users experience this to be "true", CMC will have proven its unique power compared to other communication means.

CMC – joining the crowd

A critical factor for CMC to be regarded as useful and usable is a well-balanced relation between the complexity of the system, i.e. the number of functions and procedures supported, and the "ability" of the user interface to level with the user and his expectations. Introducing CMC to a new user society will always demand a lengthy and well-coached period of training. Advanced on-line assistance may help out in these matters (Wasson & Akselsen 1992).

Studying why people fail to use CMC – why they don't join the crowd – might be interesting. Even more interesting are studies on why people quit using CMC – why the drop out of the crowd. It turns out that implementation of CMC systems, maybe more than any other computer-based system, should be need-driven. The implementation should not start "with a technology in search of an application, but with a collaborative group which has an important and enjoyable task" (Hiltz 1989). They should find CMC as a most appropriate way of communicating regularly, and they should all have easy access to the system, i.e. have their own terminal at their work-place. Minimalising the requested user-effort should be one of the main goals of the design of the CMC and its environment.

CMC – getting used to talking electronically

Exactly what is the nature of CMC, and how may CMC prove useful? And with its speed, does CMC end up as a means of increasing the information overload? Surely, the speed and power of CMC might remind us of McLuhans global village: "Ours is a brand-new world of allatonceness. "Time" has ceased, "space" has vanished. We now live in a *global* village ... a simultaneous happening. () At the high speeds of electronic communication, purely visual means of apprehending the world are no longer possible; they are just too slow to be relevant or effective. () As soon as information is acquired, it is very rapidly replaced by still newer information." (McLuhan & Fiore, 1967).[2] CMC and electronic communication has too often been presented with a most promising future, as e.g. by Tesler (1991): "In five or six years' time it will be as natural to collaborate through a network as it is to prepare a holiday feast with friends in a common kitchen".

CMC has been subject to studies of usage and effects of use ever since its introduction. The introduction of the EIES system was for instance accompanied by field experiment for assessment of the new technology (Hiltz 1978). Such field experiments may have methodological disadvantages, but they have given valuable insight into the nature of CMC and mediated communication as such.

The nature of communication through CMC differs more from the nature of other written communication than what was first expected by e-mail designers. Although being a medium for the written communication, some studies show that users of e-mail soon adopt a more verbal and informal writing style than in both traditional mail and in use of telefax. The differences may be due to the fact that typing something on a computer screen is regarded as less work, and thus as less formal, than putting some message on a sheet of paper. Or as Sproull and Kiesler (1991) puts it, social posturing and sycophancy are reduced which may be regarded advantageous. Also, a study on the comparative use of different communication media cited by Motiwalla et. al. (1992) indicated that users found face-to-face communication and telephone more appropriate for exchanging rich information. A more in-depth study by Lea (1991) shows that users themselves find CMC most like note-writing and letter-writing and least like face-to-face communication. Leas users did also rate CMC as just as asynchronous as letter-writing even though electronic transfer is that much faster. This is an interesting, if not unexpected, result. Further, "E-mailing was construed to be equally spontaneous as face-to-face and telephone conversation" (Lea 1991).

Worth noting are also a set of results from a study on computer support for

[2] *George Washington once remarked, "We haven't heard from Benj. Franklin in Paris this year. We should write him a letter." (McLuhan & Fiore, 1967)*

work groups consisting of retired employees and employees at work but eligible to retire (Eveland & Bikson 1988, Hahm & Bikson 1989). The electronically supported groups in this experiment developed structures taking advantage of the electronic media in terms of breadth of access and opportunity to participate and asynchrony increased the ability of non-collocated retired members to take an active role. More interestingly, the electronic groups maintained significantly higher degree of contact and they had considerably less communication isolation. Electronic groups did also experience significantly more involvement. Last, but not least, electronic communication did not turn out as a substitute for traditional media. The electronic groups did instead maintain higher levels of communication through all channels.

Other effects, the reduction of politeness and concern for others, are the main disadvantages with electronic communication. "In adding the memory and processing power of computers to communication technology, electronic communication does much more than speed up information flow. It loosens constraints of space, time, numbers of people, social or organisational boundaries, and information ownership on communication" (Sproull and Kiesler 1991). Not surprisingly we thus find a category called "Threats and put-downs" in McCormick and McCormicks (1992) study of the contents of undergraduates' e-mail. During a six month period more than 3 % of the messages exchanged fell into this category. Although the users in this case were a youthful, small sample of undergraduates, "flaming" is found in most e-mail environments.

Regarding "emotional quality" Leas users rate e-mail slightly towards the poor end of the dimension, whereas face-to-face communication and telephoning were seen as activities providing a "facility to express emotions" (Lea 1991). Leas users were using e-mail in their business and inter-organisational communication, therefore one would expect them to do less socialising across the network. McCormick and McCormicks users were young students, and more than half of the replies to why they used electronic mail was that they did it to socialise. For these users, CMC "enhanced intimacy instead of stifling it" but CMC was not seen as a substitute for a face-to-face support network. Conversely, of those students that did not use e-mail, less than one fifth of the answers gave "Prefer private conversations" as the reason of non-use of CMC (McCormick and McCormick 1992). Most studies of CMC-usage show that heavy users of e-mail report having more face-to-face communication than the light users. The low emotional quality of CMC have however given rise to a whole set of new grammatical constructs like e.g. :-) showing a smile (mostly used to indicate that a specific phrase is to be taken as a joke), :-0 showing that the writer is shouting at you; ;-) showing that the writer is winking at you, etc.

CMC have thus proven to establish new ways of interpersonal behaviour. And CMC have been looked at as a tool which might enhance present working situations. Electronic mail is the general-purpose CMC tool which leaves the user more in control of the situation than some of the more advanced groupware tools which forces the users to behave according to specified patterns. Which is the better, the former or the latter, is all dependent upon the context in which the communication takes place.

Learning through the network

CMC have given a new dimension to the world of distance learning. With CMC available it turns out that students in distance learning may more easily establish and maintain interpersonal networks and they are given easier and more convenient "access" to their tutors (Fjuk & Jenssen 1989, Castro 1991, Rasmussen et al 1993). It seems that CMC fills the need for a way of exchanging questions and answers, problems and suggested ways of dealing with those in an asynchronous manner. And, taking into account the situation of the distance learner who holds down a job or has a family to look after, CMC is particularly suited (Castro 1991). In this manner, CMC turns out to be a powerful addition to the pre-packaged instruction material.

CMC has been used for constructing virtual classrooms (Hsu & Hiltz 1991) and has proven to be extremely helpful in teaching collaborative skills to students. Through the virtual classroom, a set of software in combination with the EIES CMC system, students were grouped into teams carrying out role playing, event-oriented scenario games. A control group was given the same tasks to perform, but were not given access to the virtual classroom. Not only did the students with access to the virtual classroom gain better results in the games, but they did also establish a closer social network and "camaraderie" (Hsu & Hiltz 1991). Confer also the results of the field experiment with the retirees cited above (Eveland & Bikson 1988, Hahm & Bikson 1989).

In another study (Harasim 1991) it was found that "delivery of education through computer communication alters the relationship of the instructor, the students, and the course content" – students engaged in collaborative learning. The same study does however point to the importance of the interface of the CMC and the integration with other computer-based tools for handling information, including the information exchanged through CMC.

Integration, and the structuring of information, may further strengthen the position of CMC in distance education. One such system is based on the use hyper-structures as means for communicating in a local area network (Romiszowski 1990). This experiment came about after some initial experiments with plain CMC in distance learning.

The integration of CMC with access to databases have also proven as a powerful means of gathering information in the process of learning.

CMC and the health care environment

Surviving in the information society is more a question of gaining access to the most relevant information for the situation at hand than a question concerning the existence of the information. In other words, it is not that much a question of "is the information somewhere to be found?" than a question of "how do we go about finding what we want?" "Communication systems can provide a way to assess the need for medical help and avoid unnecessary visits to doctors, as well as encourage necessary ones." (Greenberger & Puffer 1989).

Most works concerning computing in health care which touch upon the concept of communication translates this into the networking of information systems (databases). The basis of these stories are always the same: the vast amount of medical information which may be relevant for problem solving, and the large number of sources of relevant information. This is true for e.g. Rennels and Shortliffe (1987) in their work on advanced computing for medicine: "The need for communication systems arises in part because it is increasingly difficult for a physician (or a biomedical investigator) to read, memorise and remember all the information needed to solve a particular problem. () Physicians have accordingly started to use literature data bases, or bibliographic retrieval systems." The same is true for O'Desky et. al. (1990) when they try to identify evolving computer technologies and further try to anticipate what effect they will have by the year 2000.

Also, we seldom find CMC considered as relevant in works on medical informatics and education (e.g. the special issue of the Methods of information in medicine, vol. 28, 1989). From this we could have concluded that CMC has no future neither in the daily work in health care, nor in the education of health care workers.

A counterexample is found in the book by Ellis (1987) where he looks at computing and applications in medical environments where a few passages briefly touches the concept of CMC and its potential: "A particular growth area is in the computer technology analogue of the postal service: electronic mail. () Whether this is between individual doctors or health centres and hospital departments, electronic mail offers much scope for breaking the tyranny of the paper-based health service. Furthermore, such facilities as bulletin boards allow doctors and other health care professionals to exchange views and information in a more immediate and informal fashion than the traditional letter in a journal." (Ellis 1987, page 28). This gives hope for CMC as a method of establishing and maintaining professional interpersonal networks. Further, Ellis finds applications for CMC in relation to the general practitioners information systems: "... what the doctor could do with is: a) more efficient channels of communication between his or her clinical environment and the outside; b) assistance in sorting the informational wheat from the chaff; and c) practical strategies for handling knowledge in a more positive manner. () Obvious applications of electronic mail in general practice include sending referral letters to out-patient departments, receiving discharge summaries direct from hospitals, the rapid reporting of pathology results, and so on." (Ellis 1987, page 139). So Ellis points directly to, for instance, the application of CMC in the distribution of information to and from patient record systems. The next step into a CMC-supported health care environment would then have to be based on an investigation into "the channels of communication between a doctors clinical environment and the outside.

The challenge

In a former paper (Danielsen 1990) we discussed the information islands and information bridges. The former is a result of implementation of information systems designed to cover the needs of the individual. The latter is the traditional method of repair through the exchange of data between information systems. For some, CMC is looked upon as a potential bridge between health care institutions. CMC and access to databases are the key applications in the MEDIS package (Engum 1991). This integration was believed to give users, i.e. physicians, a minimum of information services relevant for their needs. A first version of this package has been implemented as part of the telemedicine project and has been introduced to a number of potential users. MEDIS is built to fit with hardware currently available in the health care sector - that is PCs. That means that users should be able to access medical information services from their own workstations. A prototype version of a more advanced integrated package for information services has also been implemented. This package, known as Ratatosk, runs on Unix workstations.

In order to gather knowledge on the communicational structures within health care, a number of surveys have been undertaken. These surveys cover general practitioners offices (Danielsen 1991a, Danielsen 1991b), use of external laboratory tests for general practitioners offices (Stenvold & Karlsen 1991) and a radiology department at a specific hospital (Braa et al 1991, Engevik 1991). Some of the results of the survey at GP offices are:

- one out of two letters received are related to previous communication (letters sent or phone conversations)
- more than 90 % of letters received are stored - often electronically
- for 5 to 10 % of letters received it would have been seen as an improvement if they were received earlier
- less than half the letters sent are to a few select receivers (hospitals, labs etc.) the rest are mainly to patients
- nearly all letters sent are based on information likely to be stored electronically (patient records).

It turns out that speed is not a crucial issue if other media than ordinary mail is to be considered. Less than half of the units to be mailed had receivers which could be believed to have, or be likely to invest in, CMC technology. Although a large quantity of the information received through mail was related to earlier communicative acts, most where not as parts of longer discussions through external mail. This means that most mail received were not parts of incremental communicative processes. We could therefore conclude that CMC could have been a useful tool for handling most of the information which today is communicated through ordinary mail. The transmission speed of the medium used was of little interest, but the pre- and post-storing of the information could be a significant factor in favour of CMC. Observations and interviews did not uncover urgent needs for more group-oriented CMC tools.

A number of forecasts and "dreams" of a golden age for CMC have been given since the birth of computers. Whether these forecasts were wrong with respect to the number of people using or the frequency with which they use CMC are of less interest. What is of interest to note is that CMC has not turned out to be a key reason for investing in technology for any CMC type of user, i.e. people do not buy computers because they need to or like to use them for communicating with other people. Yet, technology for person-to-person communication is seen as one of the information technologies which will have a significant organisational impact (Straub & Wetherbe 1989).

However promising forecasts may be, we should not expect CMC to take over as the one most used general purpose communication medium in health care. CMC may strengthen its position in health care environments *if,* and only if, it proves to help out in improving productivity. As for CMC and productivity, a study (Hiltz 1988) shows that pre-use expectations correlates the strongest. That means, for CMC to "conquer" the market, potential users should state belief in CMC as the right tool. Further, the study shows that the perceived value of information provided by peers, time spent on-line, perception of problems with this mode of communication and how many users get to know while on-line are process variables playing a role in determining positive productivity outcomes. Lastly, "the social context and software differences will interact to affect the most productive applications of the system". That means that the competitiveness and the co-operation between group members have to be considered carefully. As to the simplicity of the CMC system it turns out that the easier the system is to understand initially, the more are users likely to limit what they will attempt to accomplish.

Consequently, CMC should be applied to the non-incremental communicative processes where user expectancies with respect to improved productivity are considered as positive. This would be the case for CMC integrated with computer-based systems for storing patient-oriented information. Furthermore, in the case of group-oriented CMC, e.g. conferencing systems, the emphasis on non-incremental communication should be strengthened. Health care information relevant for such systems are therefore not found in the lengthy discussions between colleagues - even if they are geographically dispersed, but in the processes where data and news are posted for a larger audience which could benefit from commenting or responding to whatever was posted, e.g. a hospital announcing changes in their procedures, etc.

The challenge of CMC is to prove itself as the highly context- and problem-specific tool in daily work routines of health care, and not to present itself as merely another general-purpose method for communicating.

References

Archer, N P. 1990. A comparison of computer conferences with face-to-face meetings for small group business decisions. *Behaviour and information technology,* 9, 307-317.

Benford, S, Turoff, M, Palme, J. 1992. An ISO standard to support asynchronous group communication. *Computer Standards & Interfaces*, 14, 363-373.

Braa, J, Monteiro, E, Vedel, E. 1991. *Kartlegging av kommunikasjonsstrukturer ved sykehus. (Qualitative description of communication for a hospital).* Oslo, Norwegian Computing Centre. (NR-report 847).

Castro, A S. 1991. Using computer-mediated communication to support the Australian adult distance learner. *Interactive learning international,* 7, 67-74.

Clausen, H. 1991. Electronic mail as a tool for the information professional. *The Electronic Library,* 9, 73-84.

Cook, D, Ridley, M. 1990.*Computer-mediated communication systems: will they catch on?* CLJ, December, 413-417.

Danielsen, T. 1990. *A possible projection of the future: potential effects of information technology in the health service.* Paper presented at the "Computers in Hospital Care" conference in Vienna, June 1990. Kjeller, Norwegian Telecom Research. (NTR-lecture TF F 9/90).

Danielsen, T. 1991a. *Kartlegging av kommunikasjonsstrukturer ved kommunale legekontor. Rapport fra Lødingen legekontor (Qualitative description of communication for general practitioners offices. A report from a study in Lødingen).* Kjeller, Norwegian Telecom Research. (NTR-report TF R 8/91).

Danielsen, T. 1991b. *Kartlegging av kommunikasjonsstrukturer ved kommunale legekontor. Rapport fra kommunelegekontoret i Nordreisa (Qualitative description of communication for general practitioners offices. A report from a study in Nordreisa).* Kjeller, Norwegian Telecom Research. (NTR-research document TF N 22/91).

Danielsen, T et al. 1991. Ratatosk - an adaptive user interface to computer-mediated communication. In: *Message handling systems and application layer communication protocols.* Schicker, P. & E. Stefferud (eds.) Amsterdam, North-Holland, 345-354.

Ellis, D. 1987. *Medical computing and applications.* Chichester, Ellis Horwood Limited, ISBN 0-7458-0093-9.

Engevik, R. 1991. *Kartlegging av kommunikasjonsstrukturer. Undersøkelse ved Røntgenavdelingen, Lovisenberg sykehus (Qualitative description of communication for hospitals. A report from a study at the radiology department of Lovisenberg).* Oslo, Diakonhjemmets høgskolesenter. (Thesis).

Engum, B. 1991. *MEDIS - Medisinsk informasjonstjeneste (MEDIS - Medical information service).* Kjeller, Norwegian Telecom Research. (NTR-report TF R 22/91).

Eveland, J D, Bikson, T K. 1988. Work group structures and computer support: a field experiment. *ACM Transaction on office information systems,* 6, 354-379.

Finnset, W. 1992. *Brukermodellering og brukermodell i Ratatosk - generelle implikasjoner og skisse til design (User modelling in Ratatosk - general implications and sketch of design).* Kjeller, Norwegian Telecom Research. (NTR-report TF R 14/92).

Fjuk, A, Jenssen, A E. 1989. *EDB-basert kommunikasjon i fjernundervisningen (CMC in distance learning).* Oslo, University of Oslo. (Thesis for the degree of cand.scient in Informatics).

Greenberger, M, Puffer, J C. 1989. Telemedicine: toward better health care for the elderly. *Journal of communication,* 39, 137-144.

Hahm, W, Bikson, T K. 1989. Retirees using email and networked computers. *International journal of technology and ageing,* 2, 113-123.

Harasim, L. 1991. Designs & tools to augment collaborative learning in computerised conferencing systems. In: *Proceedings of the 24th annual Hawaii international conference on system sciences* (HICSS), 1991, 379-385.

Hartvigsen, G. 1992. *Adapterbare og adaptive brukergrensesnitt (Adaptable and adaptive user interfaces).* Kjeller, Norwegian Telecom Research. (NTR-report TF R 2/92.)

Hiltz, S R. 1977. Communication medium. *Technological forecasting and social change,* 10, 225-238.

Hiltz, S R. 1978. Impact of a computerised conferencing system upon scientific research specialities. *Journal of research communication studies,* 1, 111-124.

Hiltz, S R. 1988. Productivity enhancement from computer-mediated communication: a systems contingency approach. *Communications of the ACM,* 31, 1438-1454.

Hiltz, S R. 1989. Computer-mediated communication systems: dropouts versus users. In: *Proceedings of the IFIP 11th World Computer Congress.* Amsterdam, North-Holland, 1989, 1089-1094. ISBN: 0-444-88015-1.

Hsu, E Y P, Hiltz, S R. 1991. Management gaming on a computer mediated conferencing system: a case of collaborative learning through computer conferencing. In: *Proceedings of the 24th annual Hawaii international conference on system sciences (HICSS),* 1991, 367-371.

Lea, M. 1991. Rationalist assumptions in cross-media comparisons of computer-mediated communication. *Behaviour & information technology,* 10, 153-172.

McCormick, N B, McCormick, J W. 1992. Computer friends and foes: content of undergraduates' electronic mail. *Computers in human behaviour,* 8, 379-405.

McLuhan, M, Fiore, Q. 1967. *The medium is the massage.* London, Penguin Books.

Motiwalla, L F, Nunamaker, J F. 1992. MAIL-MAN: A knowledge-based mail assistant for managers. *Journal of organisational computing,* 2, 131-154.

O'Desky, R I, Ball, M J, Ball, E E. 1990. Computers in health care for the 21st century. *Methods of information in medicine,* 29, 158-161.

Pankoke-Babatz, U. (ed.) 1989. *Computer based group communication. The AMIGO activity model.* London, Ellis Horwood. ISBN 0-7458-0694-5.

Quarterman, J S, Hoskins, J C. 1986. Notable computer networks. *Communications of the ACM,* 29, 932-969.

Rasmussen, T et al. 1993. PROFF: *Computer-mediated communication in Nordic research education: a social feasibility study.* Kjeller, Norwegian Telecom Research. (NTR-report TF R 7/93.)

Rennels, G D, Shortliffe, H. 1987. Advanced computing for Medicine. *Scientific American,* 257, 146-153.

Richardsen, P W, Danielsen, T. 1989. Intelligent messages or when messages come alive. In: *Network information processing systems.* Boyanov, K. & R. Angelinov (eds.) Amsterdam, North-Holland, 257-264.

Romiszowski, A J. 1990. Computer mediated communication and hypertext: building structure into distance seminars. In: *Proceedings of Pacific telecommunications: weaving the technology and social fabric.* Wedemeyer, D J. & M D. Lofstrom (eds.) Honolulu, Jan. 1990, 540-547.

Smith, H, Onions, J, Benford, S. (eds.) 1989. *Distributed group communication: the AMIGO information model.* London, Ellis Horwood. ISBN 0-7458-0741-0.

Sproull, L, Kiesler, S. 1991. Two-Level Perspective on Electronic Mail in Organisations. *Journal of Organisational Computing,* 1, 125-134.

Stenvold, L A, Karlsen, E. 1991. *Kartlegging av analyserekvisisjoner fra kommunalt helsesenter. Rapport fra en studie ved Båtsfjord helsesenter (Qualitative description of the use of external laboratory tests for general practitioners offices. A report from a study at Båtsfjord helsesenter).* Kjeller, Norwegian Telecom Research. (NTR-research document TF N 16/91.)

Straub, D W, Wetherbe, J C. 1989. Information technologies for the 1990s: an organisational impact perspective. *Communications of the ACM,* 32, 1328-1339.

Tesler, L G. 1987. Networked computing in the 1990s. *Scientific American,* September 1991, 54-61.

Wasson, B, Akselsen, S. 1992. An overview of on-line assistance: from on-line documentation to intelligent help and training. *The knowledge engineering review,* 7, 289-322.

Whittaker, S. 1991. *Towards a theory of mediated communication.* Bristol. (HP Laboratories, report No. HPL-91-78.)

Restraining and facilitating factors in the diffusion of telemedicine – An interview study

BY DEEDE GAMMON

621.39:61

1 Background

There appears to be every indication that telemedicine will spread rapidly within the Norwegian health care sector. A wide range of telemedicine applications has already been established as permanent services in several health care institutions, and more hospitals are anxious to establish their own services. The Norwegian Ministry of Health and Social Affairs has accredited telemedicine as a specialist field, and has appointed the University Hospital of Tromsø as a national competence centre devoted to furthering developments in telemedicine. If today's wide and positive (both popular and professional) press coverage continues, telemedicine will soon become a term both understood and anticipated by the general public.

In spite of these positive developments, however, there is every reason to ask: How will telemedicine fare within the Norwegian health care sector? What factors will influence the diffusion process – that is the rate and pattern of dispersion. While both technical and user group experiences are well documented, social, economic and organisational issues now come to the foreground as essential in assessing the potential diffusion of this technology. These issues now come to the foreground in our efforts to anticipate and influence the future of telemedicine within the Norwegian health care sector.

The interview material reported in this paper was our first effort in gaining a better idea of the factors which can influence the process of diffusion within the Norwegian health care sector. Subjects who have participated in, or been affected by Norwegian Telecom's telemedicine R&D activities were interviewed based on the following general question: *What factors can be expected to have restraining and facilitating effects on the diffusion of telemedicine within the health care sector?* They were asked to formulate "educated guesses" about the future of this technology as they saw it from each of their varying experiences and positions: local and central hospitals, municipality administrations, the Ministry of Health and Social Affairs, Norwegian Telecom, and the Norwegian Telecom's primary collaborators.

The goal of the study was to

- gain a preliminary overview of the range of factors which are considered relevant to the diffusion of telemedicine
- gain a basis for designing measures which can contribute to mutual adaption between telemedicine technology and health care organisations.

2 What can be gained by the diffusion of telemedicine?

Ideally, the primary factor facilitating the diffusion of telemedicine is its ability to achieve important health care goals, in addition to the goals of telemedicine-related industry. Indeed, the majority of interview subjects emphasised *all that telemedicine can help us to achieve* as the major factor which would facilitate the diffusion of telemedicine. We will not discuss the validity of this assumption (that "do-good" technologies diffuse quickly, while others do not). Here we will merely provide a summary of the reasons why Norwegian Telecom as well as the health care sector are concerned with the issue of diffusion.

The Norwegian Telecom has fairly straight forward motives for its efforts in developing and spreading telemedicine; increase the use of telecommunications and thus increase the profits. Furthermore, the health care sector provides challenges which are important to tackle in order to maintain dominance in other areas of the public and private sector.

The health care sector on its part is obligated to provide satisfactory health care to all members of society within the economic framework provided by the government. Here, the legitimacy of telemedicine will be assessed according to its ability to contribute in achieving important goals within the health care sector. The following list shows the types of goals or potential benefits which both health care personnel in the telemedicine pilot projects and interview subjects have formulated:

- More efficient and equal access to medical specialists, less waiting time
- Reduced patient travel
  - Lower transport costs
  - Fewer and shorter leaves of absence (due to rapid diagnosis)
- Reduced number of ambulatory services
  - Lower specialist travel expenses
  - Increased specialist productivity
- Improved knowledge transfer to rural physicians
  - Accreditation of distant consultation as part of specialist education
  - Better access to patient cases necessary for maintaining hospital
  - Fulfilment of criterion for specialist accreditation
- Fewer number of referrals to specialists
- Reduced turnover of medical personnel among rural health care institutions due to isolation and lack of access to wider professional network
- Avoidance of unnecessary surgery (telepathology)
- Improved quality of health care services due to better co-ordination and continuity of treatment, and better information to the patient.

To date, the projects have concentrated primarily upon assessing the technical and medical feasibility of the applications. The above types of goals or potential benefits have functioned as informal guidelines for assessing the relevance and/or feasibility of the various pilot projects. Few of the goals listed above have been operationalised and assessed systematically. This will be among the primary concerns in future studies.

3 Interview study

Twenty-nine interview subjects were selected according to two criteria: a) acquaintance with the developments of telemedicine, either directly through participation in pilot projects, or indirectly through contact and co-operation with the Telemedicine project, and b) representation of areas of interest and influence as shown in table 1.

It is worth noting that several of the interview subjects are pioneers who have invested a fair amount of energy and prestige in developing and implementing telemedicine applications. Furthermore, the study was conducted by a Norwegian Telecom research scientist also involved in this field. The questions and opinions of potential users and institutions who have *not* been involved in these developments will be important in gaining a

Table 1

Areas of interest and influence	Number interviewed
Health care institutions	
Central hospitals	
Medical personnel	5
Administrative personnel	3
Rural hospitals	
Medical personnel	3
Administrative personnel	2
Primary health care	1
Health care authorities	
Ministry of Health and Social Affairs	1
Finnmark municipality	1
Troms municipality	1
Nordland municipality	1
Rogaland municipality	1
Norwegian Telecom (NT)	
Tromsø Telecommunication Area	1
North Regional Administration	1
ComStrat	1
Research	1 (10)*
Telecom Business Communication (TBK)	1
Others	
Patient ombudsman, Nordland municipality	1
Competence Centre for IT in the Health Care Sector (KITH)	1
NovaKom A/S	1
Fearnley Data (now EdiCom)	1
Spacetec	1
Total	29

* 10 research scientists at the Telemedicine project participated in a group interview lasting approx. 45 mins. This is counted as one interview.

more complete and reliable idea of the types of factors that will influence diffusion. At this stage, however, few apart from those directly involved with telemedicine have enough knowledge and experience to have developed opinions on these issues.

The subjects were asked the following open question: *What factors can be expected to have restraining and facilitating effects on the diffusion of telemedicine within the health care sector?* While no specific subset of questions was formulated, we prompted the subjects to comment on the following types of issues; personnel, economics, organisation and technology. Within this framework the subjects were free to follow the lines of thought they found most relevant based on their varying experiences and positions.

4 Facilitating and restraining factors

The interview material is summarised into nine groups of factors which emerged as central throughout the interviews. For each group, the material is organised into facilitating and restraining factors as indicated in the shadowed boxes. Since most of these factors are self-explanatory, we have limited our comments to a) supplementing general impressions, b) areas subject to misinterpretation, and c) issues where the subjects had seriously divergent views. It should be kept in mind that several of the factors may be both facilitating and restraining, depending upon the context at hand. Furthermore, some of the factors presented are descriptive (based on the subject's own experience), while others are normative (based on the subject's opinion of how it should be).

4.1 General social and political factors

The ongoing public concern and debate about the "crisis" in the health sector creates a readiness to assess newer and more radical measures in helping solve some of the structural and economic problems in the health care sector. The majority of subjects in our study reflected this attitude towards telemedicine. Further, if broad and positive media coverage is any indication, telemedicine appears easy to understand and to "sell" to the public. As one subject put it, "This is a type of technology that the politicians will fall for." A simple, practical demonstration of telemedicine is often enough to give lay persons perspectives of how this technology can help attain important values and goals in society; equal access to generalist and specialist health care, more quality health care per krone.

The restraining factors most frequently mentioned were related to mechanisms regulating the roles and responsibilities between the different levels of public health administration. These mechanisms, often described as rigid and complex, are constantly under public assessment and debate. In most cases, telemedicine is used between institutions administered at different levels or between, for example, municipalities in different regions. Among the problems which arise in this interaction is the issue of economic compensation tied to the health care services themselves (travel, consultation fees, etc.) as well as the issue of dividing the burden of investments in telemedicine systems – an issue which is complicated by a strong trend towards decentralised goal management.

Restraining

- Rigidity and complexity of mechanisms regulating national, municipal and local activities
- Decentralised goal management of investments and measures which are in the interest of all parties
- Exaggerated focus on technology rather than goals

Facilitating

+ Social/political values and demands for:
  · Equal health care to all citizens
  · More health per krone
+ The "LEON principle" stating that health care should be provided at the level closest to the patient
+ Trends towards cross-sector collaboration
+ Public debate on changes in traditional hospital structure

4.2 Characteristics of telemedicine system solutions

The only characteristic listed under facilitating factors which was *descriptive* of today's telemedicine applications was that they had *potential* for achieving the goals referred to under section 2. Most subjects described today's telemedicine applications in terms of restraining factors. Telemedicine's overall success to date, which was acknowledged by the majority of interview subjects, is clearly *in spite of* its present characteristics.

It is therefore important to emphasise the fact that telemedicine has to date been driven by enthusiastic pioneers within a research and development (R&D) context. The technical and organisational problems experienced, both within the health care field sites and within Norwegian Telecom's own organisation, have been perceived as challenges to be tackled within this context. This will not be the case for future users who pay for the systems, and who will lack the support of enthusiastic R&D personnel. Future user groups can be expected to be more apprehensive and critical towards the changes imposed by this type of technology. Few problems need arise before telemedicine would be rejected as unsuitable.

Restraining

- Expensive equipment for limited applications
- Users (of VC studio) are dependent on organising their time around other user groups
- Causes annoying and time demanding break with existing routines (new and confusing booking procedures, running to and from VC studio)
- Vendor dependency, lack of standardisation
- Technical instability (periodically poor sound quality, break in communications lines)

Facilitating

+ Potential for achieving important goals (see "What can be gained ...")
+ Packaged telemedicine products (including technology, service, training, consultation services tied to implementation and organisation)
+ Modularity, mobility, decentralised (desktop)
+ Standardisation
+ Faster and cheaper
+ Simple additions to existing medical technology (microscope, endoscope, etc.)
+ Easy to learn and use

All of the subjects who expressed opinions on this issue emphasised the characteristics listed to the right as critical before broad diffusion of telemedicine could be expected. Several expressed high expectations towards Integrated Service Digital Network (ISDN) which will support text, data pictures and speech through a single plug in the wall. If ISDN achieves projected functionality, it can be expected to accelerate diffusion of both existing and new telemedicine applications. Flexible, cheaper, office based solutions will lower the threshold for a wide range of user groups.

In addition to necessary technical improvements, several underlined the importance of well organised routines for preparation and booking of the studios. Procedures for ensuring the presence of trained personnel in both the sender and receiver studios before and during the consultation must be established. This means new responsibilities for technical and administrative personnel, and must be figured into implementation costs and plans.

4.3 Norwegian Telecom (NT) as vendor

We received two types of general comments from health care representatives concerning NT as a vendor; confidence and satisfaction with collaboration at the local and regional levels, along with a somewhat sceptical, wait-and-see attitude towards NT's policy making at the central level. The latter had primarily to do with the tempo at which policies for pricing and organisation of the product and service apparatus were clarified.

NT's marketing and service departments have traditionally been organised around telecommunication products. In later years, these departments began organising their services around customer categories or branches. While the health care sector has been categorised as a branch during the last 2 - 3 years, it presents unique challenges which NT has only recently begun to work on in a systematic way.

NT's lack of an overall strategy for the health care sector was a major concern for the majority of subjects both inside and outside NT. This is reflected in the interview material in that the restraining factors listed on the left are predominantly descriptive of today's situation. The short term consequence of this is the time lag expected before telemedicine is available as a commercial product. The health care sector needs guidelines for both short term and long term planning and investments in telecommunications. Due to lack of workable market and service organisation, research scientists are often tied up in operative tasks for user institutions.

Active measures are being taken to correct these problems. NT is implementing an encompassing reorganisation which will have implications for NT as a competitive organisation in general, as well as its operations within the health care sector in particular. NT's marketing strategy for the health care sector will be completed medio 1993. Mechanisms for ensuring rapid transfer of R&D products to the market organisation are also being implemented.

Restraining

- Lack of strategy towards the health care sector
- Premature release of new products, promising more than can be delivered
- Pricing policy
  - Late and unclear signals
  - Lack of overall structure
- Fragmented, uncoordinated products and services as well as the roles and policies of the various departments and suppliers
- Lack of knowledge about the health care sector
- "Paranoid" of being accused of cross-subsidising
- Poor contact and utilisation of NT's own R&D department

Facilitating

+ Health care customers should have few (preferably one) point of access to NT
+ Overall positive contact/co-operation with NT at the local level
+ Large and dependable organisation
+ Greater freedom/flexibility in commercialising telemedical products
+ Competition
+ Knowledge of and strategy towards health care sector
+ Active use of introduction prices

4.4 Economic and legal issues

The difficulties imposed by today's system of transferring funds between national insurance authorities, county municipalities and local authorities, was mentioned by a majority of the subjects.

For example, the national insurance authorities cover travel expenses (including transportation, diet and hotel) for ambulating specialists, as well as travel expenses for patients. Thus, they save money by replacing ambulatory services and/or patient travel with distant consultations. It is the county municipalities and local authorities, however, who bear the burden of investing in telemedicine. Further, the central hospitals (who supply specialists for distant consultation) find their incomes dwindle due to the reduced number of guest patients from neighbouring municipalities.

Today's regulations state that specialist fees are contingent upon the *physical* presence of the specialist. In many cases, the local physician is together with the patient while the specialist performs a diagnostic examination via the telenet. According to the regulations, this is not a specialist consultation, and is thus ineligible for coverage of specialist fees.

Several legal issues must also be resolved. For example, who is legally responsible for the patient – the distant specialist, or the local physician who is together with the patient? Those who expressed opinions on this issue were unanimous in support of today's regulation which places full legal responsibility with the specialist. Nevertheless, situations may arise where the issue is not definite. The Ministry of Health and Social Affairs is collaborating with the University Hospital of Tromsø in order to clarify legal and economic guidelines for distant consultations.

Restraining

- Complicated and rigid system of transferring funds between parties
- The benefits of telemedicine are not enjoyed by those who bear the costs
- Unclarified legal questions

Facilitating

+ Settlement based on actual cost/benefits for each party
+ Each party benefits while increasing the amount of health per krone (which some believe possible)

4.5 Trends in health care

A series of factors – here called trends in health care – were expected to influence the diffusion of telemedicine.

Several expressed a belief that the strong and inevitable trend towards sub-specialisation in the health care sector would be among the most pervasive facilitative factors in diffusion. (An example of sub-specialisation is an ear-nose-throat specialist who specialises in ears.) Sub-specialisation implies fewer specialists per subgroup – the number of which – is growing. This, in addition to the expense of highly specialised personnel and the accompanying medical technology, compounds the trend towards centralisation of care delivery, and thus also adds to the challenges of attaining the goal of equal access to quality health care.

It is argued that this is among the major reasons why the effects of hard handed cost containing measures are "eaten up" by higher quality standards. Theoretically, the quality services is heightened, but it is increasingly difficult to attain the goal of equal access to quality health care.

This, combined with an increase in expense and quality standards caused by rapid diffusion of high-tech medical technology will – according to some – necessitate telemedicine.

Field trials involving a distant specialist who examines the patient together with the patient's general practitioner indicate that the GP – after a series of these specialist consultations – is able to diagnose and treat a greater number of patients

Restraining

- The filling of available positions for physicians
- "Protectionism" between health care professions
- Shutting down of rural hospitals

Facilitating

+ Increased sub-specialisation
+ "High-tech" medical equipment which heightens quality norms for – and expenses of – health care services
+ Ideals of a "wholeness" approach to patient treatment
+ Upgrading of general practitioner's position as the patient's primary advocate
+ Network concepts in organising health professionals
+ Increasing need for extended education and training
+ Increasing need and demand for more flexible collaboration between hospitals and the general practitioner (e.g. before hospital admittance, and under rehabilitation)

(some estimate up to 50 %) before referring them to the specialist. The endoscopic trials will investigate this question more systematically over a period of a year. This type of knowledge transfer to rural physicians is important in achieving the LEON-principle which states that treatment should be provided at the lowest possible level of the medical professional "hierarchy", and as close as possible to the patient's local environment. The ideals of continuity in the treatment program and consideration of the patient's total life situation – both previous to and following hospital admittance – was also mentioned as a factor expected to facilitate diffusion of telemedicine.

4.6 Centralisation and decentralisation

The Norwegian society in general – the health care sector included – highly values the maintenance of decentralised public services. Thus, measures which facilitate decentralisation are viewed as positive while those involving centralisation of services are considered necessary evils forced upon health care by lack of resources and demands for rationalisation.

Several subjects were concerned that telemedicine would reduce pressure for recruiting physicians in rural areas, while others were convinced that telemedicine would increase the attraction of rural positions and thus support recruitment.

Telemedicine confuses some of the traditional ways of defining centralisation/decentralisation. Future considerations of this issue should build upon a renewed assessment of the advantages and disadvantages of each. It is important to keep in mind that it is policy and practical use which will determine the consequences of telemedicine along this dimension, not the technology itself.

4.7 Physicians

Several subjects expressed the opinion that doctors – particularly specialists – would be the single most influential group effecting the diffusion process. As one put it, telemedicine won't have a chance if doctors don't like it – regardless of how supportive hospital management or politicians might be. Likewise, if doctors respond to telemedicine in a positive way, its success is assured.

Restraining

- Telemedicine has centralising effects

Facilitating

\+ Telemedicine has decentralising effects

Although none of the doctors interviewed acknowledged the factors on the left as their personal attitudes or interests, they were confident that these factors would play an important role in acceptance and diffusion. In discussing this issue, it is useful to distinguish between two groups:

- Expertise delivering physicians
- Expertise receiving physicians

The first group consists primarily of specialists at central hospitals who either have ambulated (and enjoyed supplementary income) or have contributed significantly to hospital income by treating guest patients from neighbouring municipalities. By carrying out distant consultations, they experience a reduction in income both for themselves and their hospital. Furthermore, some physicians have shown a negative attitude towards allowing *receiving* (primarily rural) hospitals to participate (via video conference) in medical meetings, while they are positive towards participating in similar meetings when they themselves are recipients.

The specialists interviewed were asked why they were positive towards distant consultations. These expressed an interest in new and exciting technology, as well as emphasising the advantages for the patient and the health care system in general. Again, these specialists are among the pioneers in telemedicine developments in Norway, and it is doubtful that their attitudes and motives are representative of their colleagues as a whole.

The second group – expertise receiving physicians – has to date consisted of physicians at rural hospitals and institutions in the primary health care. Through distant consultations with specialists, these physicians have built up a new degree of competence and thus an increased ability to screen, diagnose and treat patients locally. Our subjects assumed that this would promote acceptance and diffusion of telemedicine.

Restraining

- Preference for travel due to
  . supplementary income
  . relaxation, break from daily hospital "drudge"
- Protection of own profession
- "Surgeon syndrome" – conservatism towards all other technologies than one's own
- Unsatisfactory work environment
- Unsatisfactory telemedicine ergonomics
- Fear of "surveillance"

Facilitating

\+ Tired of – or outright refusal – to perform ambulatory services
\+ Prefer using travel time for productive or rewarding work
\+ Place importance on family and leisure time
\+ Increase in number of female physicians
\+ Idealists – seek the well being of patients and the health care sector

Rural physician access to a wider professional network, as well as accreditation of specialist consultations as part of the continuing education curriculum, was also emphasised as an important incentive. Among the potentially restraining factors for this group's acceptance of telemedicine was the fear of surveillance or control from central experts, as well as concern for diluting their role as general practitioners.

The effects of telemedicine on the physician's work environment was underlined as a factor influencing acceptance and diffusion. Some were concerned with the consequences of working long periods with video or screen based consultations. The importance of organising the number and type of consultations in order to allow a varied work schedule was emphasised.

The effects of telemedicine on the physician's work environment will also interplay with other environmental factors. Physicians working in environments characterised by overload and stress will most likely perceive implementation of telemedicine as an additional burden.

This, coupled with telemedicine's replacement of ambulatory services (i.e. "fleeing the drudge of the hospital ward") as well as a possible fall in income, will doubtlessly muster resistance – if not outright sabotage – towards implementation of telemedicine. Development of reasonable incentives as well as measures for integrating telemedicine into the work environment will be important concerns.

4.8 Organisational culture

Organisational culture reflects a given organisational entity's values and norms and thus also its ability to accept and grasp change. It affects the way implementation processes are dealt with, and thus also the rate at which the organisation is able to enjoy the benefits of innovations such as telemedicine.

The factors grouped and listed below are limited to user organisations – in our case hospitals, hospital wards, primary health institutions, etc. Cultural issues related to other relevant organisational entities (e.g. the health care sector as a whole, vendor organisations such as Norwegian Telecom) are scattered throughout the interview material under other factor group headings.

Restraining

- Job insecurity due to hard-handed rationalisation policies
- Poor experiences with similar technology implementation ("more technology dumped over our heads")
- Exaggerated focus upon short-term savings
- Personnel generally worn out
- Implementation designed solely by individual enthusiasts, thus making the organisation vulnerable to turnover
- Lack of supportive management
- Lack of incentives for change

Facilitating

+ Management support and view that technical implementation implies organisational development
+ Oriented towards goals and common vision
+ Open information flow and personnel participation in implementation process
+ Tolerance for enthusiasts/those who take risks
+ Value status of modernised operations
+ Hope for "saving angel" (e.g. small rural hospitals threatened with closing down)

Both the facilitative and restraining factors listed above can be recognised as having general relevance for any encompassing technical innovation and implementation process. Diffusion of telemedicine implies far more than purchase and installation of telemedicine equipment and communication services. It implies new roles and ways of interacting between professional personnel, as well as new tasks for technical and administrative personnel. The decisive role physicians are expected to play in the diffusion of telemedicine (see previous section) is in part due to the powerful effect they can have as opinion leaders – and thus also the organisational culture – within health care institutions.

4.9 Cost/benefit assessments

Several subjects expressed the view that the time is right for initiating cost/benefit assessments. It is assumed that this type of documentation is necessary in order to obtain "diffusion facilitative decisions" from both health care authorities and telemedicine related industry. Few, however, offered specific views as to how such studies should be performed, or if and how the results would be used.

Some were concerned that cost/benefit studies could do more harm than good if the studies weren't conducted properly. There was an awareness that it is methodologically difficult – as well as touchy – to find meaningful measurements (i.e. "price tags") for the wide range of elements involved in providing health care services (e.g. to rural areas) – with and without telemedicine.

Restraining

- Assessment made too early or too late
- Assessment based upon one party's point of view
- Tendency to focus upon short timespan
- Tendency to ignore qualitative factors which are difficult to measure

Facilitating

+ Used to demonstrate need for restructuring reimbursement system
+ Used as basis for tailoring distant consultation fees

5 Summary discussion

The interview material gives an idea of what persons with telemedicine experience believe will be important restraining and facilitating factors in the diffusion of telemedicine. The following factors which emerged have been discussed:

- General social and political factors
- Characteristics of telemedicine system solutions
- Norwegian Telecom (NT) as vendor
- Economic and legal issues
- Trends in health care
- Centralisation and decentralisation
- Physicians
- Organisational culture
- Cost/benefit analysis.

Obviously, these types of factors cannot be summed up to any given conclusion or prediction as to how telemedicine will fare within the Norwegian health care sector. The general impression we are left with is that there is as much working for – as there is against – the diffusion of telemedicine. We can, however, attempt to summarise general impressions of the major factors at play.

The most obvious facilitative factor which emerged from the interviews is the *belief* that telemedicine will benefit both patients and the health care sector. The majority of subjects – many of which are decision makers – repeatedly emphasised the potential benefits of telemedicine as the major facilitative factors for the diffusion of telemedicine (see "What can be gained ..."). Although we cannot assume that "do-good" technologies diffuse more effectively than other technologies, we can assume that the opinions of those interviewed will play an important role in the diffusion process in Norway. The subjects belong to a network of health care, administrative and political professions which – compared to many countries – are relatively personalised and transparent in Norway. Those who saw potentially negative consequences of telemedicine, suggested ways to avoid these consequences. These suggestions included technical, organisational, economic and legal measures, each of which were believed feasible.

The major overall restraining factor which emerged throughout the interviews is the *complexity* of the two parties playing decisive roles in diffusion: Norwegian health care sector, and the Norwegian Telecom.

The health care sector – as the recipient and user of telemedicine – must implement a series of economic and organisational measures in order to achieve the projected benefits of telemedicine. Telemedicine – the use of which often breaches traditional organisational boundaries – will be hampered by the rigid roles and division of responsibility between national, county and local authorities. Characteristic of large, intricately rule-governed systems is a built-in sluggishness in the types of decision making processes necessary in designing and implementing new policies conducive to utilisation of innovations like telemedicine. Similarly, the process of developing and implementing medical and technical standards into an integrated "health network" (emphasised as decisive for telemedicine's success) can be characterised by tug of war between a wide range of interest groups within the health care system. These characteristics of the health care sector are particularly challenging for vendors of information and communication technology. As opposed to many other business customers, there is no co-ordinating IT-strategy to which vendors can tailor their product and marketing strategy.

Norwegian Telecom (NT) is the major locomotive for telemedicine related industry in Norway. NT itself is undergoing encompassing changes as it moves from a monopoly to a competitive role. The splitting of products and services into ONP and competitive groups, and the emergence of a series of small companies as part of the liberalisation process, makes NT appear like "a troll with many heads" as one subject described it. Few have a clear idea as to how these different entities will collaborate in marketing telemedicine within the health care sector. Customer demand for packaged solutions (including all necessary equipment, communication, training, installation, service) and fewest possible (preferably one) vendor relationship, is a formidable challenge for telemedicine related industry. Another important challenge is development of the type of competence needed within the marketing organisation in order to anticipate and meet the diverse needs within the health care sector.

We contend that a great number of the restraining factors which emerged during this study can be understood in light of the complex nature of the health care sector's and NT's organisation – and their consequent ability to grasp the acknowledged potentials provided by telemedicine. The technical problems described by our subjects as restraining would appear to be the least concern. These types of problems receive continuous attention and are being solved at a rapid rate. As we try to anticipate the diffusion of telemedicine, it is important to keep in mind that telemedicine is not a single *product* (e.g. PC, telephone, or new type of medicine) to a single *customer category* (e.g. Norwegian households where one or two persons decide purchase and routines for use). Realisation of telemedicine's potential benefits demands fundamental change within both receiver and vendor organisations.

Background references

Andersen, K A. 1992. *Videokonferanser i helsevesenet: erfaringer fra to års bruk i Troms og Finnmark.* Kjeller, Norwegian Telecom Research. (TF paper N3/92.)

Foss, G, Jøsendal, O. 1991. *Fjernkonsultasjoner.* Kjeller, Norwegian Telecom Research. (TF report R9/91.)

Gammon, D. 1991. *Evaluering av teleradiologiforsøket v/RST - TMS; aksept, bruk og diffusjon.* Kjeller, Norwegian Telecom Research. (TF report R67/91.)

Valvåg, H et al. 1992. *Applied Telemedicine.* Tromsø, Tromsø teleområde. Kjeller, Norwegian Telecom Research.

Liberalisering av telekommunikasjonsområdet

AV KJELL STORDAHL

621.39.05

Dette foredraget ble presentert på konferansen "Liberalisering gir store muligheter innen telekommunikasjon", holdt på Oslo Plaza 12 og 13 mai 1992.

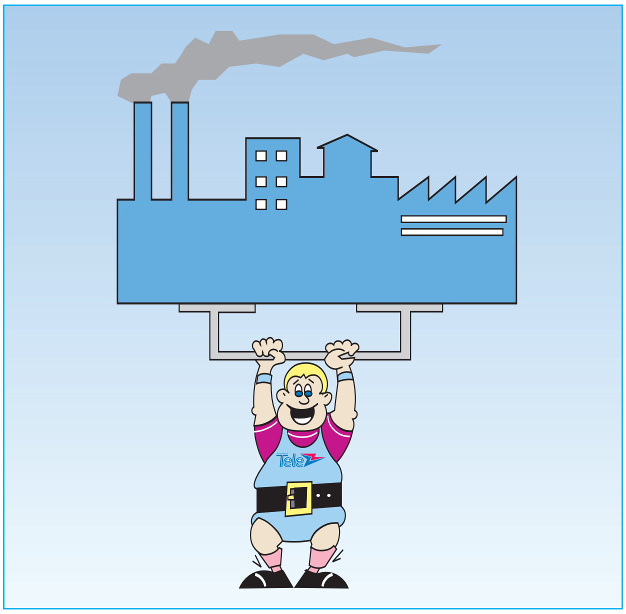
1 Innledning

På mange måter kan en si at utviklingen innen telekommunikasjonsmarkedet er *eksponentiell.* Tidligere var det få tjenester, høye priser og lav servicegrad. I 70-årene og fram til midt på 80-tallet var det lange ventelister på primærtjenesten telefon. Det var også en framkommelighet i nettet som etter vår målestokk nå var meget dårlig.

Fra kun å ha telegraf-, telefon- og telekstjenesten har markedet nærmest eksplodert. Bare i tilknytning til de digitale sentralene kan det tilbys mange titalls tilleggstjenester. Datatjenestene er i sterk utvikling. Dette gjelder blant annet overføring ved ulike hastigheter og ved bruk av en rekke ulike nettkonsepter. På overordnet nivå i nettet (intelligent nett-tjeneste) vil det være mulig å definere nye tjenester uten å foreta hardware/software utskifting i de underliggende digitale sentralene.

I tillegg har vi fått inntoget av de verdiøkende tjenestene – de tjenestene som mange snakket om for 5 år siden, men som få visste hva var. Teletorg og andre informasjonstjenester er eksempler på disse.

Fra midt på 80-tallet foretok Televerket store investeringer i fornyelse av telenettet basert på ny teknologi. Bruk av mikroelektronikk, digitale og optiske komponenter har ført til færre feil, bedre framkommelighet, større fleksibilitet, flere tjenester og større kapasitet i telenettet. Den moderniseringen som Televerket til nå har foretatt, har vært meget lønnsom og den har gitt Televerket og norsk næringsliv et godt utgangspunkt.



Figur 2.1 Televerket understøtter norsk industri

Situasjonen nå er at Televerket gir god service samtidig som prisene den siste tiden har falt til et nivå som er lavere enn europeisk gjennomsnitt. Fra å være fokusert på servicegrad, må Televerket nå i stor grad orientere seg mot markedstilpasning, lønnsomhet og konkurranse.

Tidligere var det mange private telefonselskaper i Norge. De dekket riktignok ulike geografiske områder. Disse ble kjøpt opp og Televerket har i mange år (altfor mange) vært i en monopolsituasjon. Dereguleringsog liberaliseringsprosessen både i og utenfor Europa fører nå i et akselererende tempo til full konkurranse på alle områder innen telekommunikasjon. Dette gjelder også det norske markedet.

Et forbehold kan være at det ikke vil være konkurranse på ulønnsomme samfunnspålagte oppgaver. Dersom Televerket ikke får nødvendig handlefrihet for å møte den tiltagende konkurransen, og som følge av dette taper vesentlige markedsandeler, vil finansiering og driftstilskudd over statsbudsjettet være eneste måte å dekke opp ulønnsomme samfunnspålagte oppgaver og etter hvert også deler av sin primærvirksomhet.

2 Televerket og norsk næringsliv

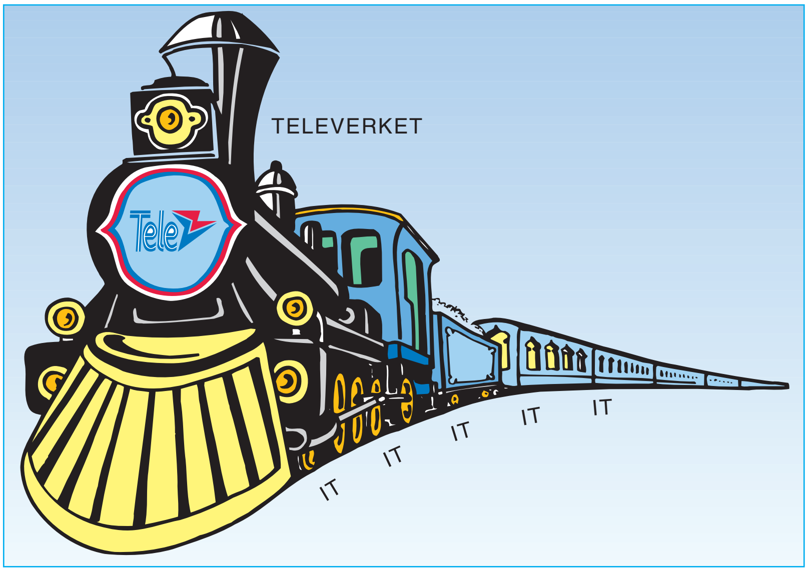
Det må kunne slås fast at Televerket og norsk næringsliv er gjensidig avhengige av hverandre.

Televerket tar sikte på fortsatt å være den sentrale nettoperatøren for norsk næringsliv. Televerket skal være en kunderettet telekommunikasjonsbedrift som med framtidsrettet teknologi og medarbeidere med høy kompetanse skal dekke kunders behov for tale-, tekst-, data- og bildekommunikasjon med tilhørende tjenester. Tjenestene skal ha høy kvalitet og et konkurransedyktig prisnivå.

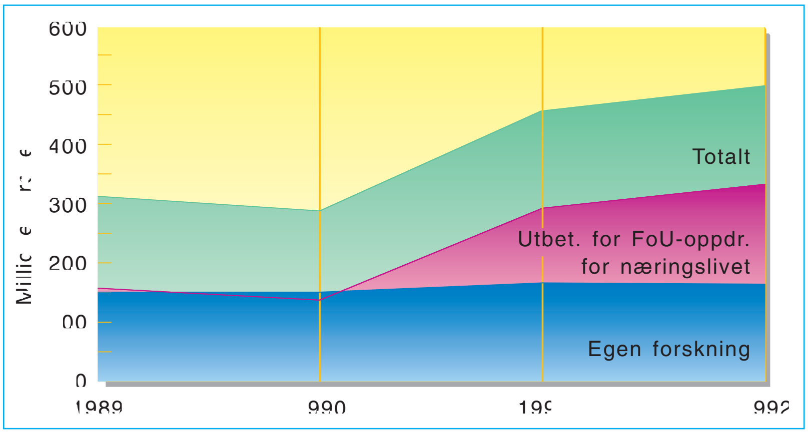
Televerkets hovedmål:

- *Televerket skal være den ledende norske leverandør av telekommunikasjoner i 90-årenes konkurransemarked.*
- *Televerket skal tilpasse sin teknologi til markedet.*
- *Televerket skal sammen med datterselskaper og andre samarbeidspartnere dekke kundenes behov for totalløsninger.*
- *Televerket skal fortsatt prioritere effektiviserings- og rasjonaliseringsarbeidet for å få kostnadene og dermed prisen på tjenestene ytterligere ned.*
- *Prisene skal senkes/holdes under europeisk gjennomsnitt.*
- *Televerket skal være drivkraften i utviklingen av nye lønnsomme forretningsområder i telemarkedet, og bidra til et slagkraftig norsk industri- og forskningsmiljø innen telekommunikasjon.*

En konsekvens av disse hovedmålene er at næringslivet skal tilbys gode, moderne og konkurransedyktige teletjenester. Det er allerede nå lagt vekt på å holde et lavt prisnivå på internasjonal trafikk. Ikke minst er dette viktig for norsk næringsliv som har større avstand til markedene i Europa enn land flest.



Figur 2.2 Televerket – et norsk IT-lokomotiv



Figur 2.3 Forskning og utvikling i Televerket

2.1 Televerkets verdiforedling

I tillegg til å være en moderne nettleverandør for formidling av trafikk, legger også Televerket gjennom sitt moderne nett forholdene til rette for verdiforedling. Televerket har et teknologisk avansert nett, med høy digitaliseringsgrad og "intelligens" som muliggjør en rekke nye tjenester. Mange nye applikasjoner kan utnytte nettets muligheter og verdiøkende tjenester etableres i nettets grensesnitt. Dette skaper mange nye muligheter for næringslivet.

Televerket betraktes nå som et *IT-lokomotiv for norsk industri*.

Bedriften er svært stor i norsk målestokk og har en stor kjerne av dyktige spesialister med meget høy kompetanse. Mange av de beste spesialistene går over til norsk industri etter at Televerket gjennom mange år har sørget for opplæring – ja, også i en rekke tilfeller bekostet den formelle høyere grads utdanning. Dette er en verdiskaping som er viktig for hele det norske samfunn.

2.2 Televerket understøtter norsk industri

Televerket tilgodeser norsk industri ved bestilling av utstyr. Eksempel på dette er stornummerkontraktene av sentralutstyr der Televerket ivaretok nasjonale interesser ved krav til produksjon av deler av utstyret på hjemmemarkedet. EFs innkjøpsdirektiver (kapittel 9) kan endre dette forholdet, men utgangspunktet vil være å preferere nasjonale interesser i den grad det er mulig.

Det samme gjelder for bestilling av ulike nettkomponenter. Televerkets årlige investeringsbudsjett er på rundt 3 milliarder kroner.

I tillegg benyttes en del konsulenthjelp og "know-how" fra industri og konsulentfirmaer.

Televerket finansierer norsk teleforskning og utvikling gjennom sitt eget forskningsinstitutt som i løpet av de siste årene har hatt en sterk vekst. Figur 2.3 viser Televerkets investering i egen forskning og utvikling og i F&U-oppdrag til næringslivet de siste årene.

Forskningsinstituttet deltar aktivt i internasjonalt samarbeid og avdelingen har vært toneangivende internasjonalt blant annet innen mobil- og satellittkommunikasjon. Egne forskningsresultater gir Televerket innpass i internasjonale miljøer og tilgang til ny teknologi på et meget tidlig tidspunkt. Særlig er det europeiske samarbeidet blitt svært målrettet og omfattende.

Forskning- og utviklingsarbeidet gir kimer til norsk industriproduksjon. Et eksempel på et lovende utviklingssamarbeid er bildetelefonen Vision – et samarbeidsprosjekt mellom forskningsinstituttet og Tandberg Telecom a/s.

For øvrig bestiller forskningsinstituttet forskningsoppdrag og utviklingskontrakter for betydelige beløp av bedrifter og institutter. Hensikten er blant annet å legge grunnlaget for mange arbeidsplasser innen informasjonsteknologi og for å utvikle teknologi for å frambringe eksportprodukter.

Televerket bygger nå opp en internasjonal enhet – Norwegian Telecom International (NTI) – som skal ha det markedsmessige, tekniske og økonomiske ansvar for alle typer internasjonale teletjenester som tale, data, video, svitsjede og faste samband i så vel jordnett som over satellitt.

Gjennom eget og samarbeidspartneres internasjonale nett skal NTI følge kundene over hele kloden. I tillegg satser NTI på konsulenttjenester innen ulike telekommunikasjonsområder. Eksempler er assistanse ved nettutbygging på Kola og i Litauen. Ved planlegging av slik nettutbygging vil Televerket kunne være en døråpner for norsk industri i utlandet – det vil si å kunne initiere åpning av nye markeder.

Televerket og norsk næringsliv er altså gjensidig avhengige av hverandre.

2.3 Trusler og utfordringer

Hvilke trusler ser vi framover mot århundreskiftet? Et worst-case scenario vil være at en sammenslutning (en eller flere) av

	Utgående samtaleminutter Millioner	Minutter pr. innbygger
Belgia	731	74
Canada	565	22
Danmark	362	70
England*	1729	31
Finland	186	37
Frankrike	1921	34
Irland	75	21
Island	118	24
Japan	764	6
Korea (sør)	188	4
Malaysia	80	4
Nederland	905	60
Norge	281	66
Portugal	126	12
Spania	611	16
Sveits	1356	202
Tyrkia	159	3
USA	5265	21

Kilde
The Global Telecommunications Traffic Report 99
*The Global Telecommunications Traffic Repo 990

Figur 2.4 Utgående internasjonal trafikk fra ulike land

større internasjonale nettoperatører går inn på det norske markedet og *skummer fløten*. Dette er også beskrevet som spekkhoggervirksomhet. I verste fall betyr dette at Televerket kun står tilbake med de ulønnsomme forretningsområdene. Dette vil i så fall få meget betydelige konsekvenser også for norsk industri. Internasjonale nettoperatører er profittbevisste. Dette kan føre til at prisene for bruk av telenettet igjen heves til et høyere nivå.

Norge ligger i utkanten av Europa. En av suksessfaktorene for norsk næringsliv er at kommunikasjonen og kommunikasjonsløsningene mot utlandet er optimale enten det gjelder transport eller telekommunikasjon. Et sterkt og konkurransedyktig Televerk vil understøtte de framtidige utfordringer og muligheter.

2.4 Norge og utlandet

Ser vi på Norges samlede produksjon av varer og tjenester (BNP), utgjorde eksporten ca. 44 % av totalvolumet i 1990 (kilde: Nasjonalregnskapet), hvorav Europa mottar ca 90 %. Til sammenlikning eksporterer Tyskland, Sverige og USA henholdsvis 37, 33 og 13 prosent av sin samlede produksjon av varer og tjenester (kilde: NHO).

Norge har altså en økonomi som er svært avhengig av handelsmessig samkvem med utlandet. EØS-avtalen og en framtidig GATT-avtale gir utsikter til forsterking av dette bildet.

Vår åpne økonomi gir seg utslag i betydelig telefontrafikk til og fra utlandet. I 1990 ble det registrert ca. 281 millioner utgående samtaleminutter fra Norge, og nesten like mye, 276 millioner inngående samtaleminutter. Fordelt på våre 4.3 millioner innbyggere gir dette ca. 66 samtaleminutter pr. innbygger, eller ca. 134 samtaleminutter pr. hovedabonnement. *Disse tallene er forholdsvis høye i internasjonal sammenheng:* Om lag 50 % høyere enn i Frankrike og England, og tre ganger høyere enn i USA. Se figur 2.4.

2.5 Det nasjonale marked

Sammenliknet med andre land har altså Norge et relativt lite hjemmemarked. Likevel er Norge, med sin desentraliserte struktur i offentlig og privat sektor, et land hvor lange avstander spiller en stor rolle.

På slutten av 80-tallet har store deler av norsk næringsliv økt sin produktivitet gjennom oppsigelser og økt integrasjon. Spesielt har industrien stor fokus på kostnadseffektiviserende tiltak. Kjeder har vokst fram bl.a. innen detaljhandel og hotell- og restaurantnæringen. Telekommunikasjon er et av områdene som gir mulighet til slike tiltak.

Store kunder med spredt geografisk struktur utgjør derfor et segment av tele-markedet som vil friste eventuelle inntrengere.

Store bedrifter utgjør allerede en konkurransefaktor i form av egne nett, leide linjer og bruk av multipleksere som komprimerer tale og data, slik at ledig kapasitet i digitale samband utnyttes på en effektiv måte.

Eventuelle framtidige allianser mellom bedrifter/kjeder, terminalleverandører og nettleverandører vil kunne oppnå markedsposisjon.

3 EFs direktiver

Dagens næringsliv blir stadig mer internasjonalt. Telekommunikasjon er en forutsetning for et levedyktig næringsliv. Dessuten er salg av data/teleteknisk utstyr i sterk vekst. Veksten har kommet i Japan, Sør-Korea og USA, mens Europa (EF) har blitt en stor netto-importør av tele/data-teknisk utstyr. Det er bekymring for at amerikanske og japanske aktører skal bli dominerende innen telekommunikasjonsmarkedet i Europa. Eksempelvis har AT&T og SPRINT etablert seg i Europa. De er kapitalsterke nok til å bli reelle konkurrenter til teleadministrasjonene.

Denne utviklingen har medført at EF har prioritert telekommunikasjon som et satsingsområde.

3.1 EFs telekommunikasjonspolicy

Prosessen ble startet opp i 1984 da EF-kommisjonen presenterte strategiske aksjonslinjer for telekommunikasjonspolicy for EF. Hovedmålsettingen var full utnyttelse av det store potensialet som et enhetlig europeisk marked innebærer. Det skal utvikles en sterk telekommunikasjonssektor i Europa slik at det europeiske næringslivet tilbys effektive teletjenester, og det skal etableres et åpent felles telekommunikasjonsmarked preget av konkurranse.

Aksjonsprogrammet, fra 1984 og fram til i dag, har konsentrert seg om følgende områder:

- Koordinering av framtidig utvikling av telekommunikasjon
- Etablering av et felles marked for terminaler, utstyr og tjenester
- Etablering av felles forskningsprogram
- Styrking av moderne telekommunikasjonstjenester og -nett i mindre utviklede områder
- Adoptering av felles holdninger til internasjonale spørsmål.

I 1987 ble den såkalte "Grønnboken" utgitt. Her ble det fastslått at televerkene inntil videre tillates å beholde monopol på basistjenesten telefon, men at alle ikke-basistjenester skal være åpne for konkurranse. Det skal utarbeides standarder for infrastruktur og tjenester. Forvaltningsdelen skal skilles fra nettoperatøren. Typegodkjenning skal være eneste begrensning for tilkobling og bruk av terminalutstyr. Kryss-subsidiering av konkurranseutsatt virksomhet fra monopolvirksomhet tillates ikke.

3.2 Rekommandasjoner

For å gjennomføre dette aksjonsprogrammet, er det utarbeidet en rekke reguleringer, direktiver og rekommandasjoner. Eksempler på rekommandasjoner er:

- Direktiv om en felles teknisk spesifikasjon av MAC/packet-familien av standarder for direkte satellitt-kringkasting
- Rekommandasjon/MOU om koordinert introduksjon av ISDN
- Rekommandasjon/MOU om koordinert introduksjon av GSM
- TEDIS: Program for introduksjon av et kommunikasjonsnett for elektronisk utveksling av data for handel
- Rekommandasjoner om koordinering av et pan-europeisk landbasert personsøkersystem.

3.3 EFs direktiver

I 1988 ble "Terminal directive" utarbeidet. Deretter kom "Service directive" og i 1990 "Open Network Provision" (ONP). Disse direktivene legger forholdene til rette for dereguleringen av telekommunikasjonsmarkedene så langt. Hovedpunktene er:

i) Akseptering av at telekommunikasjons-administrasjoner/offentlige (PTO) fortsatt skal ha eksklusive og spesielle rettigheter i å tilby og drifte offentlig nettinfrastruktur

ii) Akseptering av at PTO fortsatt skal kunne ha eksklusive og spesielle rettigheter i å tilby offentlig telefon-tjeneste

iii) Fri konkurranse på alle andre tjenester, spesielt VØT

iv) Strenge krav til å følge standarder (ETSI) for nettinfrastruktur og viktige tjenester som tilbys av PTO og andre tjenesteleverandører for å oppnå full samtrafikk innen EF

v) Fri konkurranse på terminalutstyr innen og mellom EF-land

vi) Utskilling av forvaltnings- og nettoperatørvirksomhet

vii) Klare direktiver om åpne og ikke-diskriminerende betingelser for aksess til offentlig nett (ONP) pålegges nettoperatørene for å sikre et fritt tjenestemarked. ONP-direktivene stiller krav til tekniske grensesnitt, leverings- og brukerbetingelser og takstprinsipper. Direktivene gjøres gjeldende for faste samband, telefoni, pakkesvitsjede nett, ISDN, IN-tjenester, tjenester i kabelnett og bredbåndsnett.

I EF er det arbeidet mye for å få fram et ONP på faste samband. Det foreligger nå et utkast som er foreløpig godkjent av ministerrådet for en siste høring i Parlamentet. Ministerrådet forventes å gi en endelig godkjenning innen sommerferien. Televerkets tilbud for leide samband og digitaltjenesten vil være tilpasset dette utkastet.

Både i Japan og USA skjer det en tilsvarende utvikling mot et standardisert åpent grensesnitt. I Japan kalles det OND (Open Network Doctrine), mens det i USA går under navnet ONA (Open Network Architecture). Det satses altså i tre av verdens tyngdepunkter på å legge forholdene bedre til rette for telekommunikasjonsindustri og for næringslivet som brukere.

ONP-direktivene er utviklet for å standardisere og harmonisere tilgangen til telenettene i Europa. Det tillates for andre operatørselskaper å gå inn og leie kapasitet av televerkene og videreselge kapasitet til andre brukere. Det skal være tilgang til den eksisterende infrastruktur på like, ikke-diskriminerende, kostnadsbaserte betingelser for enhver nettoperatør.

De nasjonale nettoperatørene står tilbake med basistjenesten telefoni samt infrastrukturen hvor også eneretten er i ferd med å smuldre opp. Når tiden er moden, vil EF også komme med direktiver på dette punkt.

3.4 Direktiver for satellittkommunikasjon

EF arbeider nå med "Satellite Green Paper". Hensikten er å endre de reguleringsmessige forhold på satellittsiden gjennom et sett med direktiver. De nye direktivene kan oppsummeres ved:

- Full liberalisering av satellittjordstasjoner
- Fri adgang til satellittkapasitet under forutsetning av et tillatelsessystem som skal sikre sær- og enerettene som er forenlig med EF-retten
- Full kommersiell frihet for tilbydere av satellittkapasitet, herunder direkte tilbud av satellittkapasitet til tjenestetilbydere og kunder, under visse betingelser
- Harmonisering så langt det er mulig for å lette tilbudet av bruken av Europa-dekkende tjenester.

4 Framtidig konkurransesituasjon

4.1 Internasjonalisering

Det foregår en sterk internasjonalisering av næringslivet. Det snakkes om bedrifter uten nasjonalitet, med hele verden som sitt marked. Telekommunikasjon er i sterk vekst og den blir stadig viktigere for næringslivet både for å få mer effektiv kommunikasjon og for å spare kostnader.

Tyngdepunktene USA, Japan og Europa deregulerer sine telemarkeder. Offentlige nettoperatører privatiseres og fristilles. Konkurransen skal gi bedre og billigere tilbud til brukerne.

Markedet for teletjenester blir stadig mindre oversiktlig med hensyn til hvem som kan tilby hva. Nye nettoperatører og tjenestetilbydere etablerer seg. Næringslivets globalisering resulterer i at forskjellige lands televerk trenger inn på hverandres en gang beskyttede markeder i et forsøk på å følge sine kunder. Selv infrastrukturen til de ulike televerk er under press.

4.2 Allianse mellom nettoperatører

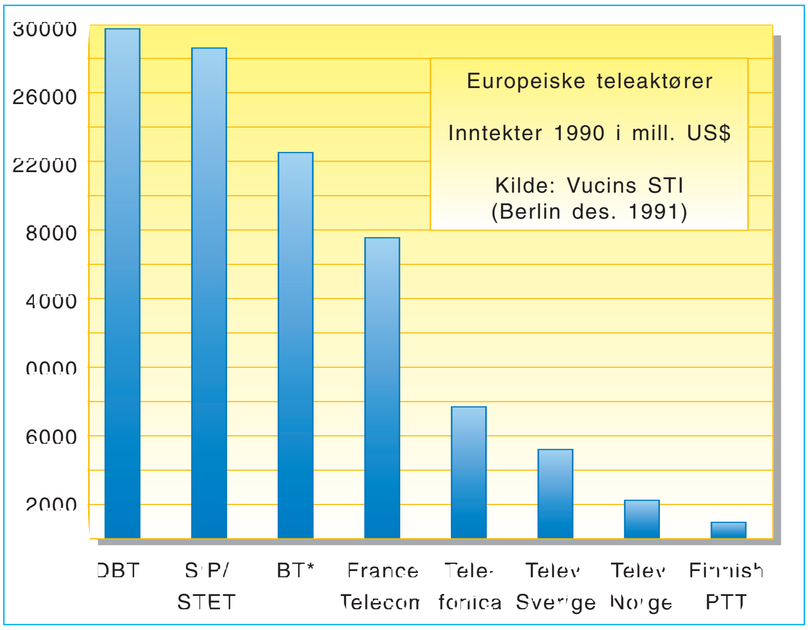
Nettoperatørene inngår allianser og de kjøper opp nasjonale og internasjonale nett. Eksempler på dette er:

- British Telecom kjøpte i 1989 TYMNET fra flyfabrikken McDonnell Douglas. Nettet har noder i 48 europeiske byer og kontorer skal åpnes i alle europeiske hovedsteder. British Telecom forhandler med IBM om overtakelse av deres interne datanett – i tilfelle vil British Telecom være representert i 154 byer bare i Europa.
- Den svenske Kinnevikgruppen holder på med å bygge opp et pan-europeisk digitalt høyhastighetsnett med hovednoder i en rekke europeiske byer.
- British Telecom har sluttet avtale med japanske KDD og amerikanske MCI. Sammen skal de tilby internasjonale kunder tilgang til data, tale, telefax og meldingstjenester integrert i høyhastighets digitale forbindelser.
- Cable & Wireless har et verdensomspennende nett. Mercury er C&Ws operatør i Europa og tilbyr europeiske kunder tilgang til nettet.
- British Telecom, Deutsche Bundespost Telekom og japanske NTT har dannet et konsortium, Pathfinder, som skal tilby totalløsninger for internasjonal kommunikasjon.
- NTT, AT&T og Ameritech åpner for samtrafikk mellom sine digitale, internasjonale nett.
- De europeiske jernbaneselskapene samarbeider nå i et prosjekt (HERMES) der de skal knytte sammen alle de nasjonale selskapenes fibernett i et pan-europeisk nett.
- Daimler-Benz holder på å bygge opp sitt private interaktive satellittnett basert på VSAT der også tale kommer til å bli overført.
- General Electric inngikk i mai 1989 en femårig avtale med AT&T, British Telecom og France Telecom om bygging av et internasjonalt nett for 75 000 medarbeidere i en rekke land.

Televerkene vil møte konkurranse ved at det bygges opp helt parallelle nett uavhengig av televerkenes nett og ved at utleid kapasitet (leide samband) benyttes i televerkenes infrastruktur.

På disse nettene vil det gå både tale og data og dessuten høykapasitetsoverføringer. Den store veksten i telekommunikasjonsmarkedet med høye fortjenestemarginer vil stimulere nye aktører til å gå inn.

Store kunder har på grunn av manglende, eventuelt for dyre, tilbud fra teleoperatørene, etablert egne private løsninger. Først er



Figur 4.1

tjenestene brukt internt; senere kan tjenestene tilbys til andre bedrifter (eksternt).

Som vist ved eksempler over, posisjonerer også televerkene seg i denne konkurransen ved å inngå strategiske allianser med andre aktører.

I Sverige etableres nå Tele 2 i full konkurranse med det svenske televerket. Tunge eiere som Kinnevik og Cable & Wireless står bak Tele 2.

Internasjonale selskaper som MCI/Infonett, IBM, Sprint, GEIS, British Telecom, AT&T, Deutsche Bundespost Telekom, France Telecom og andre har nå salgskontorer i opptil 80 byer på verdensbasis. Målsettingen for disse internasjonale teleggigantene vil være å konkurrere på datakommunikasjon og senere på internasjonal trafikk. I figur 4.1 er det vist en oversikt over inntektene til europeiske nettoperatører. Som figuren viser, har vårt Televerk liten omsetning i forhold til de største.

4.3 Konkurranse på telekommunikasjonsområdet

Viktige konkurransefaktorer i kommunikasjonsområdet vil være:

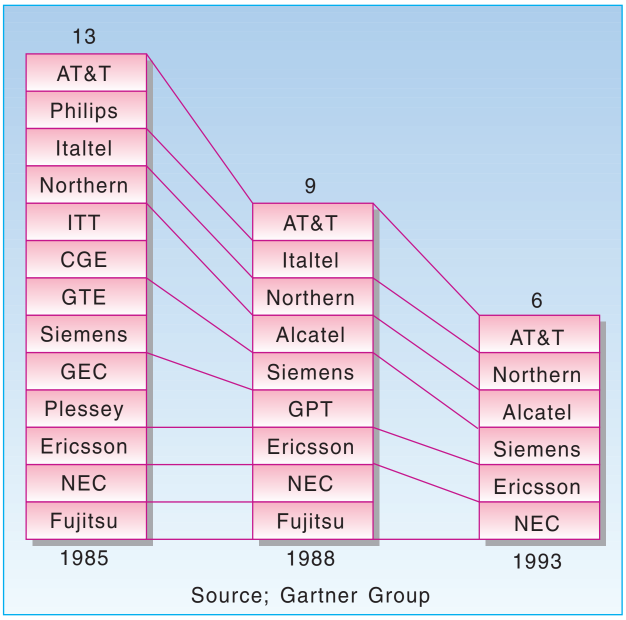
- service, opplæring, konsulenttjenester
- driftsstøttesystemer, nettovervåking
- totalløsninger.

Det konkurreres om:

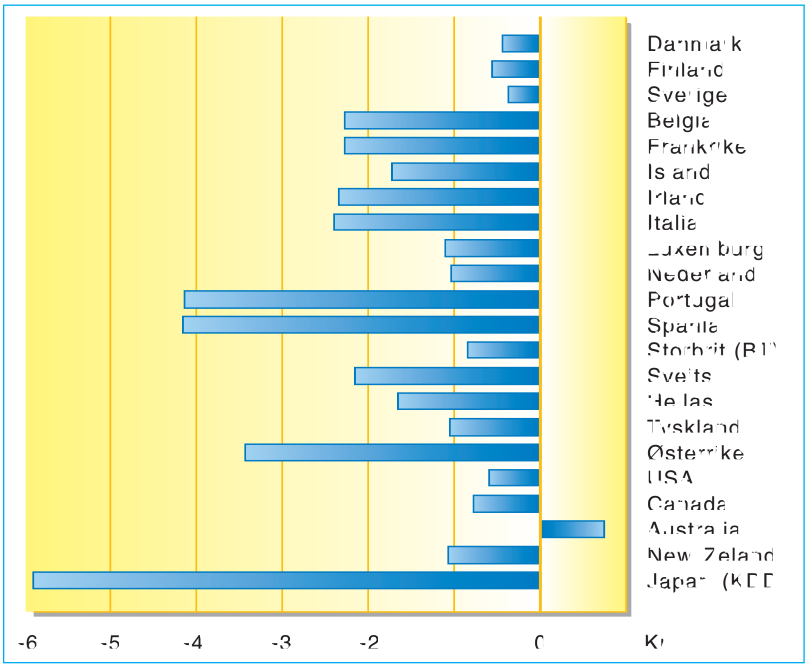
- nettoverføring
- verdiøkende tjenester.

De verdiøkende tjenestene kan deles i:

- informasjonstjenester
- prosesserings- og bestillingstjenester
- meldingstjenester



Figur 4.2 Leverandører av offentlige sentraler



Figur 5.1 Prisdifferanse for ett minutts forretningstrafikk (uten mva.) mellom Norge og andre OECD-land i høytrafikktid. Januar 1992 (i samsvar med Tarifica jan.92)

- kunderettede drifts- og vedlikeholdstjenester
- intelligente nett-tjenester.

Det europeiske markedet for verdiøkende tjenester for 1992 er anslått til å være verdt 4 - 7 milliarder US dollar og verdensmarkedet omlag 25 milliarder US dollar.

4.4 Konkurranse blant leverandører

90-årene vil bli forandringenes decennium for teleadministrasjonene. På samme måte var 80-årene forandringenes decennium for de store utstyrsleverandørene. Tidligere var de nasjonale hovedleverandørene direkte eller indirekte subsidiert. Dette er etter hvert redusert. I figur 4.2 er vist oversikt over antall leverandører av sentraler fra 1985 til 1993. Utviklingen av sentralsystemer krever store utviklingskostnader og lang utviklingstid på grunn av kompleksiteten til sentralsystemene. Dette har ført til en konsentrasjon i færre leverandører – faktisk en halvering i løpet av 8 år.

5 Priser på Televerkets tjenester

Gode prisstrategier vil være en nøkkelfaktor for suksess. Betydningen av riktige prisstrategier for Televerket kan ikke overvurderes. Prisene skal være et virkemiddel for produkter/tjenester i ulike markeder. Det bør være frihet til å kunne differensiere prisene for tjenester mot ulike markedssegmenter i større grad og mer dynamisk enn det som til nå har vært tilfelle. Televerket ønsker å få samme anledning som sine konkurrenter til å fastsette, differensiere og strukturere sine priser for konkurransetjenester.

Det er svært viktig at Televerket får den nødvendige handlefrihet til å utarbeide og gjennomføre prisstrategier i et konkurranseutsatt marked. Telemeldingen gir ikke gode nok svar på dette. Det kan lett tenkes en konflikt der politiske myndigheter eksempelvis foreslår en betydelig og rask nedbetaling av statslånet, mens Televerket ønsker en markant nedsettelse av prisene. Under Stortingsbehandlingen av Telemeldingen kom det klart fram at representantene som deltok i debattene, støttet videre prisreduksjoner.

5.1 Fastsetting av priser

Omlag 60 % av Televerkets inntektsside vedtas av Stortinget. Fra 1993 er dette delegert til Samferdselsdepartementet. Endringer av de øvrige priser fastsettes i prinsippet av Televerket uten framlegg til Stortinget. I dag er prisene satt ut fra den betraktning at prisene skal dekke driftskostnadene og gi en normal forrentning på kapitalen, samtidig som Televerket skal ivareta sosiale oppgaver og samfunnsforpliktelser.

I et konkurransemarked vil prisene bli bestemt av markedet på en helt annen måte. Markedsprisen vil reflektere kostnadsnivået til den mest effektive aktør i markedet. Prisene vil i stor grad være gitt, og det er kostnadene som må tilpasses.

Konkurranse fra utlandet vil gjøre det umulig å beholde et særnorsk prisnivå for teletjenester. Dette vil i første omgang gjelde priser for trafikk til og fra utlandet.

Televerket forbereder seg nå på å møte denne konkurransen.

5.2 Priser på internasjonal trafikk

Figur 5.1 viser en sammenligning mellom priser for internasjonal telefontrafikk mellom Norge og andre OECD land. Det er tatt utgangspunkt i en samtale på 1 minutt ut fra gjeldende takster i januar 1992. Den viser at for *forretningstrafikk (uten MVA) er det kun billigere å ringe fra Australia til Norge enn omvendt.Til alle øvrige land er det billigere å ringe fra Norge.*

For boligtrafikk der merverdi er inkludert, er det kun billigere å ringe fra Australia og New Zealand.

5.3 Priser for forretningsabonnement

Figur 5.2 viser forholdet mellom telepris og konsumprisindeksen fra 1984 og fram til 1992. Som det framgår, har konsumprisindeksen økt med over 50 %, mens teleprisen har avtatt med over 15 %. Fra 1989 til 1992 har teleprisen nominelt avtatt med over 20 %. I samme periode har det vært en inflasjon på 13 %. For forretningsabonnenter har det i samme periode vært en nominell nedgang på 40 %. *For forretningsabonnenter i Nord-Norge har den nominelle nedgangen vært enda høyere på grunn av oppheving av fjerntaksten internt i Nord-Norge. Høyeste innenlands takst der er nærtakst.*(De prosentvise angivelsene er gjennomsnitts-betraktninger basert på en bestemt "pakke" av telefonbruk for en typisk forretningskunde.)

I figur 5.3 er det vist en sammenligning av kostnaden for et forretningsabonnement i ulike land. Kostnaden er basert på fastavgift samt pris på et gitt trafikkmønster gjennom året. Dette er priser fra 1991 og de er eksklusiv merverdiavgift. Figuren viser at Norge nå har priser under et europeisk gjennomsnittsnivå. De nordiske land og Nederland har et lavere prisnivå enn Norge, for øvrig har de øvrige (vest-) europeiske land et høyere prisnivå.

For boligabonnenter er situasjonen en annen. Her ligger Norge på et høyt prisnivå, faktisk nest høyest i Europa. Dette er imidlertid ikke for øyeblikket noe konkurranseutsatt marked. Kostnaden for en boligabonnent er også her basert på fastavgift og et gitt trafikkmønster gjennom året. Det er viktig å merke seg at det ut fra kostnadshensyn *ikke er rom for særlige reduksjoner på fastavgiften* som skal dekke de abonnentavhengige kostnadene ved utbygging av nett-tilknytning.

Televerket har nå i flere år redusert takstene på langdistansetrafikk. Dette gjelder både takstene på nær/fjerntakst og på utenlandstrafikken. Dette er gjort både for å redusere de overføringer som er gjort av den gruppen som ringer mye langdistansetrafikk (vesentlig forretningsabonnenter) slik at det som abonnentene betaler for telefonbruken står mer i forhold til de kostnader som Televerket har.

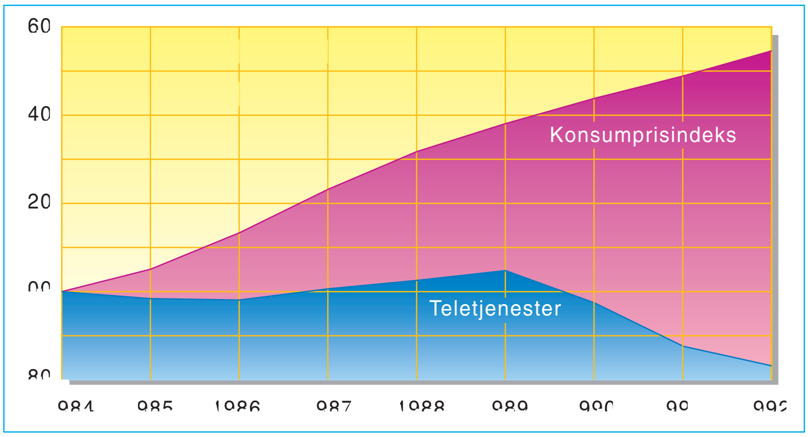
Televerket vil fortsatt differensiere framtidige takstnedsettelser blant annet ved å skille på reduksjon i fastavgift, reduksjon i lokaltakst og reduksjon i nær/fjerntakst. Reduksjon skal skje etter markeds- og kostnadsvurderinger.

Figur 5.4 viser nominelle kostnader for mobiltelefontjenesten basert på priser fra 1992. Dette er en tjeneste som i høy grad brukes av næringslivet. Det er tatt hensyn til andelen av en etableringspris samt abonnementspris og et gitt trafikkmønster over året. Kostnadene er eksklusiv merverdiavgift. Av figuren ses det at Norge ligger under et gjennomsnittlig europeisk prisnivå og blant annet foran Sverige.

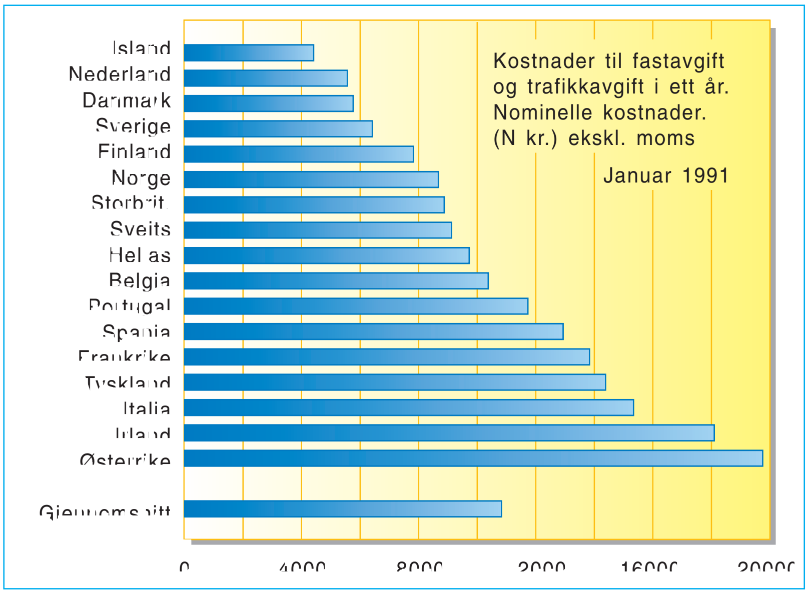
Tjenesten er pr i dag godt utbygd i de nordiske land, noe som gjør at næringslivet i Norge her har en strategisk fordel sammenlignet med øvrige land i Europa.

5.4 Prisstrategier og handlefrihet

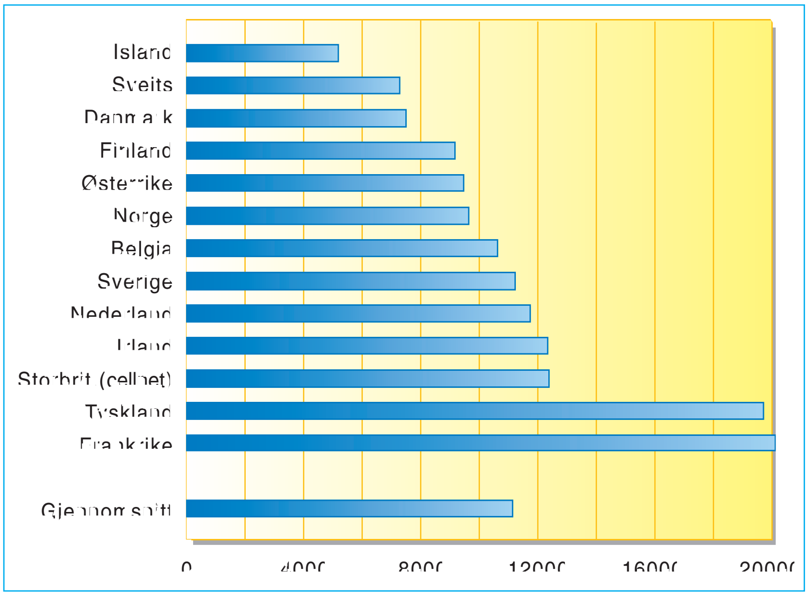
Riktig prising av leide samband er svært viktig for Televerket. Leide samband er grunnpilaren i konkurrerende nett. Fra 1/1-93



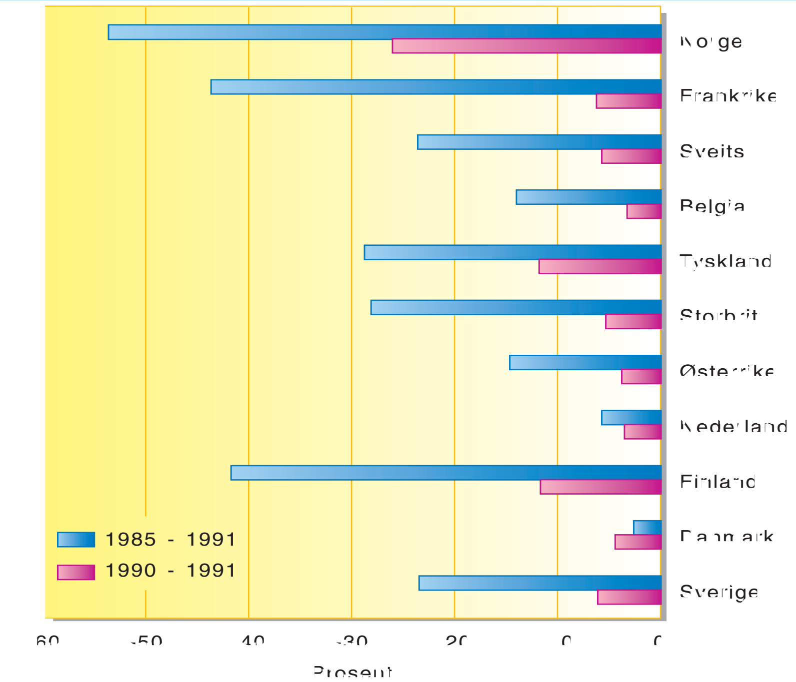
Figur 5.2 Forholdet mellom telepriser og konsumprisindeks (1984 = 100)



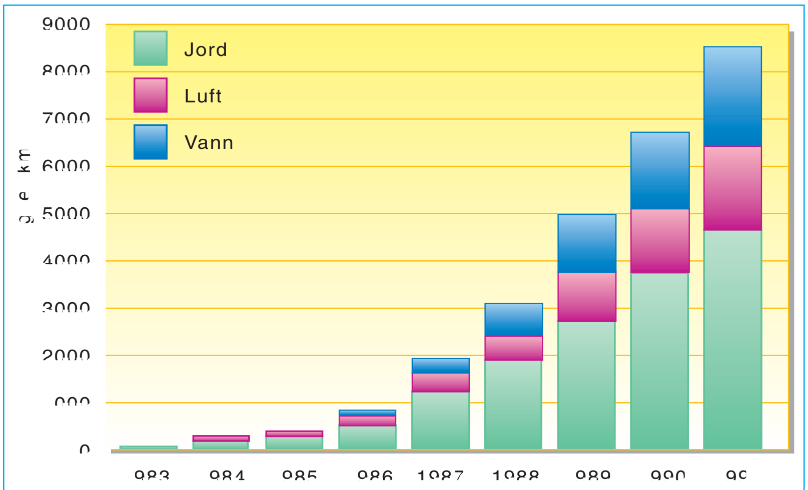
Figur 5.3 Telefon – forretningsabonnement



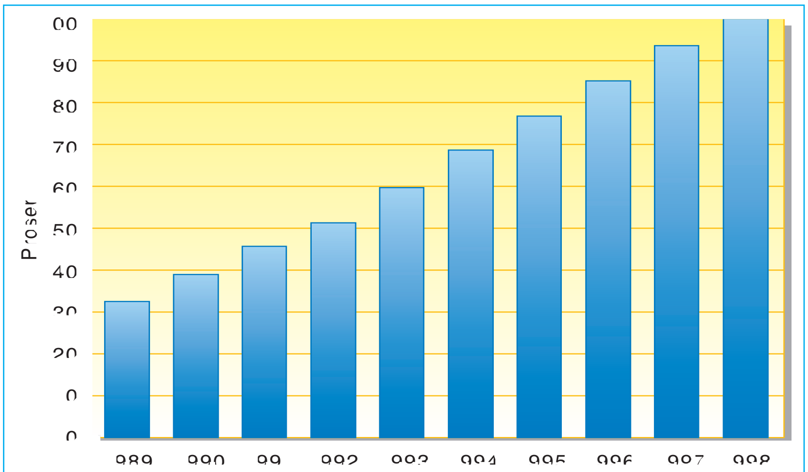
Figur 5.4 Mobiltelefonkostnader eks. mva. (basert på takster januar 1992)



Figur 5.5 Prosentvise endringer i telefonkostnader i faste priser for forretnings-abonnenter



Figur 6.1 Fiberkabel i telenettet



Figur 6.2 Digitalisering

vil enhver aktør kunne kjøpe kapasitet i Televerkets nett og samtidig selge deler av denne kapasiteten videre. Det er dette som kalles bypass. Slik utleie vil øke den kommende konkurransen.

Det vil i hovedsak være videresalg over lengre avstander som utgjør den største trusselen. Det er her flere bedrifter kan ha de samme overføringsbehovene. Etablering av konkurrerende nett vil være avhengig av Televerkets prisstrategier. Her betyr også forholdet mellom lokaltakst, nærtakst og fjerntakst mye.

Kunder i et konkurrerende nett må betale lokaltakst til Televerket når de ringer utenom eget nett. I noen tilfelle der eksterne bedrifter benytter dette nettet kan det være aktuelt å betale to ganger lokaltakst i tillegg til en form for avgift til den konkurrerende nettoperatør. *Televerket er nå i ferd med å utjevne forskjellen mellom lokaltakst, nærtakst og fjerntakst for å redusere overføringene. Dette gjør det også mindre lønnsomt å bypasse Televerkets svitsjede nett Det er derfor viktig for Televerket å kunne differensiere sine priser.*

Figur 5.5 viser ulike europeiske lands prosentvise nedsettelse av telefonkostnadene (relatert til nominelle priser) i løpet av de siste årene. Konklusjonen er at Televerket her kommer best ut. Televerket har i løpet av de siste årene satt ned prisene radikalt og planene videre er at prisene hvert år fra 1993 til 1995 skal reduseres nominelt med minst 7 %.

Televerket ønsker den samme handlefrihet som sine konkurrenter til å fastsette, differensiere og strukturere sine priser på "konkurransetjenester". Televerket må akseptere at myndighetene fastsetter rammer for prisutviklingen for "enerettstjenester", eventuelt grunnleggende krav til prisstrukturen. Innen de rammer som er fastsatt, vil Televerket ha fullmakt til å fastsette prisene.

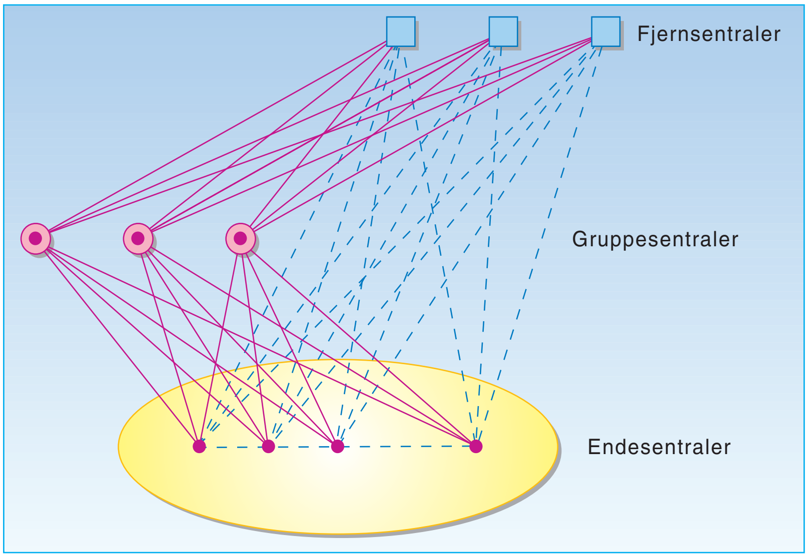
6 Televerket

6.1 Utbygging av telenettet

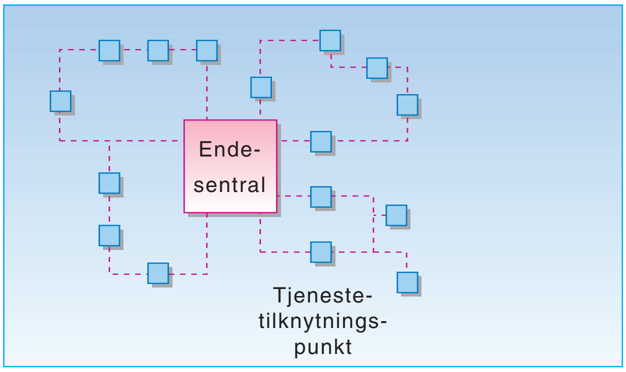
Televerket har fra siste halvdel av 80-tallet hatt en meget positiv utvikling. Det er foretatt investeringer i ny teknologi gjennom bruk av mikroelektronikk, digitale og optiske nettkomponenter som har modernisert telenettet. Dette skaper grunnlag for en mer fleksibel, mer funksjonell og sikrere videreutbygging av nettet. Moderniseringen så langt har vært meget lønnsom og den har gitt Televerket og norsk næringsliv et godt utgangspunkt både med hensyn til de topp moderne tjenester som nå tilbys og de markant fallende priser på teletjenester.

Figur 6.1 viser utbyggingen med optiske fibrer. Nettet med optiske fibrer danner en strategisk kjerne for Televerket fordi det er meget enkelt – uten for store investeringer å oppgradere kapasiteten i dette nettet. Samtidig går feilfrekvensen dramatisk ned.

Det er også foretatt en omfattende modernisering gjennom utskifting til digitale sentraler i telenettet. På overordnet nivå innføres nå "intelligent nett-tjeneste" som vil definere nye tjenester uten å foreta hardware/software utskiftinger i underliggende sentraler. De digitale sentralene tilbyr en rekke tilleggstjenester. Figur 6.2 viser den raske digitaliseringstakten i det norske telenettet.



Figur 6.3 Eksempel på nettstruktur med alternative trafikkveier



Figur 6.4 Eksempel på fiberringer i abonnentnettet

Det foretas store investeringer i nettet for å gjøre det sikrere – blant annet gjennom en reserveveipolicy – som sørger for fysisk atskilte traseer mellom ulike sentraler. I tillegg er det foretatt beslutninger på utbygging av ekstra kapasitet i sentralene ved lastdeling i tilfelle noen sentraler faller ut.

Figur 6.3 viser eksempel på denne nettstrukturen. I figur 6.4 vises eksempel på fiberringer i abonnentnettet nærmest kundene der det er to aktuelle traseer til sentralen i tilfelle feil. I de største byene vil denne strukturen bli bygget ut nå. Dette vil øke sikkerheten til kundene ytterligere.

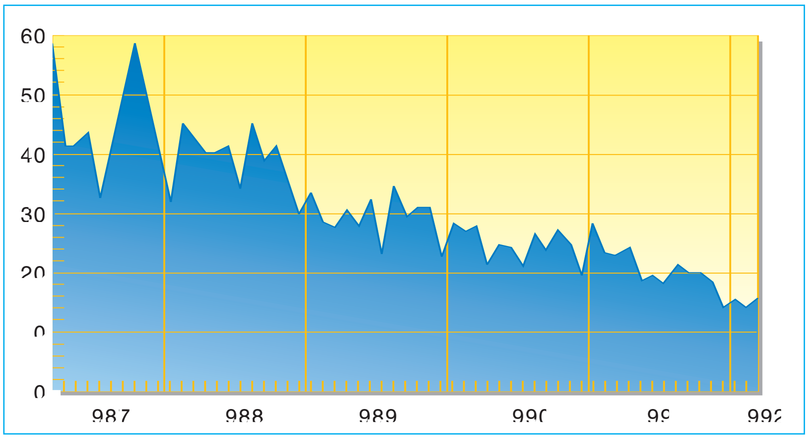
De investeringene som gjøres for å høyne sikkerheten i nettet, er i hovedsak gjort ut fra behov fra forretningsabonnenter. Boligabonnenter vil ikke ha slike krav til sikkerhet. Dette betyr derfor at boligabonnentene her gjennom betaling av sine tellerskritt finansierer deler av sikkerheten i nettet for forretningsabonnentene.På den annen side tjener Televerket forholdsvis lite på boligabonnenter på grunn av deres lavere trafikk.

Televerket er i ferd med å få et telenett med godt utbygd infrastruktur og høy pålitelighet.

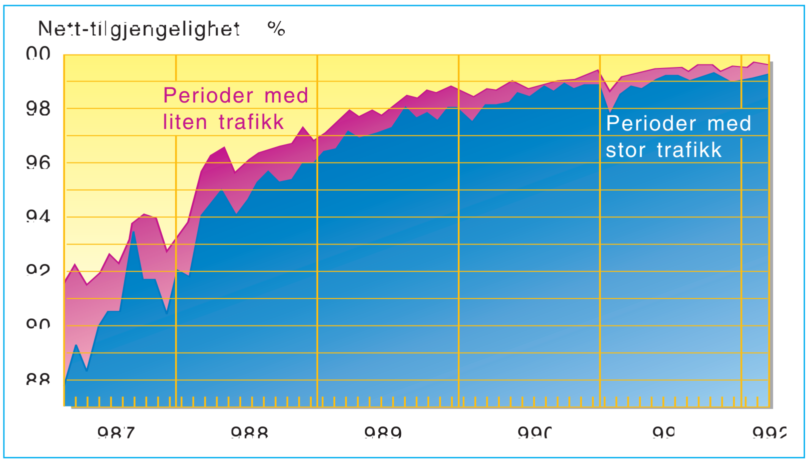
Figur 6.5 viser antall feilmeldinger pr 100 abonnenter pr år for Oslo teledistrikt. For 1992 er målet på landsbasis for antall feil pr 100 abonnenter på 15. Det ses av figur 6.5 at antall feilmeldinger har gått dramatisk ned i løpet av en femårs periode.

I figur 6.6 vises framkommeligheten i telenettet i Oslo distrikt. Også her ser vi at det i løpet av de siste årene har vært en markant økning – fra tidligere hvor framkommelighetsgraden ikke var akseptabel. Televerket har nå som satsingsområde å øke framkommeligheten mot utlandet.

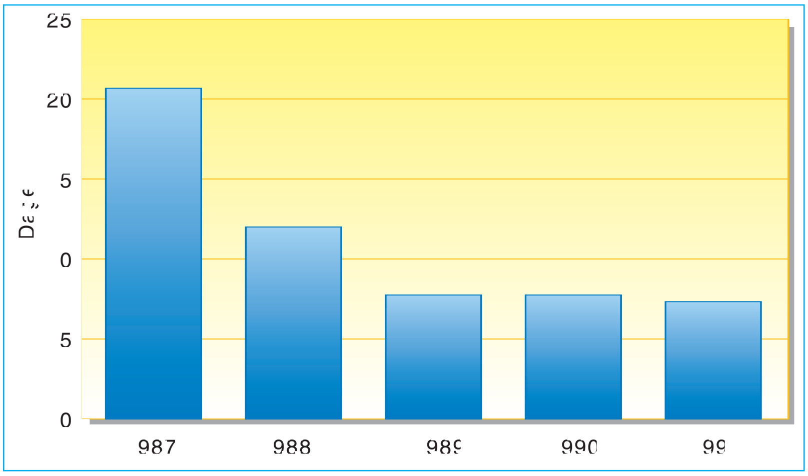
I figur 6.7 er leveringstiden for telefon vist. Leveringstiden er nå på et meget akseptabelt nivå. Ser vi tilbake, er det ikke lengre enn 6 år siden de siste ventelistene på telefon ble avviklet.



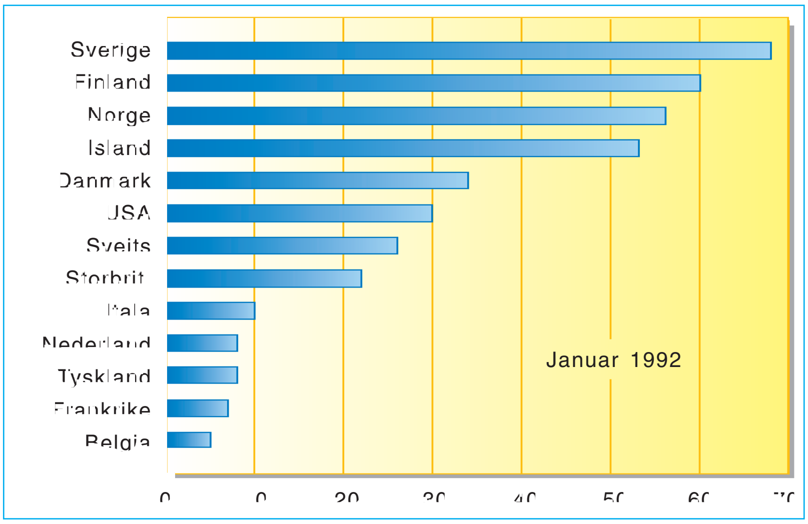
Figur 6.5 Antall kundemeldte feil pr 100 tilknutninger i Region Oslo. Utvikling januar 1987 - mars 1992



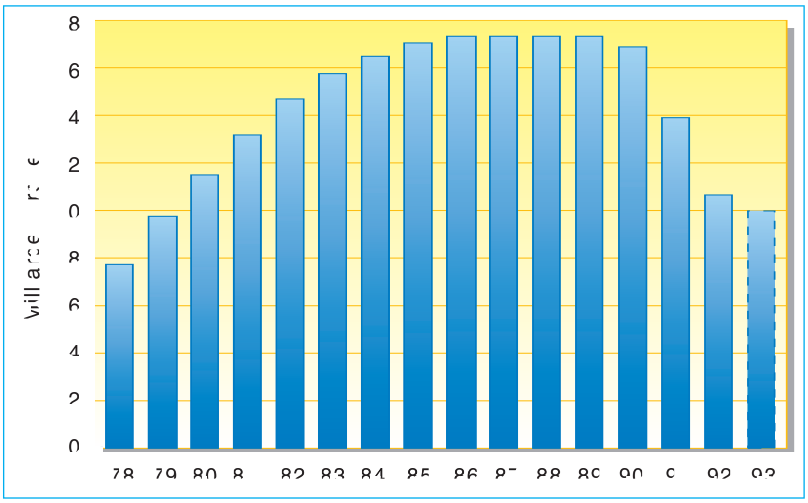
Figur 6.6 Framkommelighet i telefonnettet i Region Oslo. Utvikling januar 1987 - mars 1992



Figur 6.7 Servicestatistikk. Leveringstider for ikke tidsbestilte oppdrag, telefon



Figur 6.8 Mobiltelefoner pr 1000 innbyggere



Figur 6.9 Televerkets gjeld til staten. Restlån pr 31.12.91

Norge har i dag et ISDN pilotnett i drift. Televerket legger nå opp til at ISDN etter hvert settes i drift på kommersiell basis i Norge.

Televerket har landsdekkende datanett. Både det linjesvitsjede datanett, datex, og det pakkesvitsjede datanett, datapak, har god kapasitet og meget høy tilgjengelighet. Det er også rask leveringstid på tjenestene. Datapaktjenesten har samtrafikk med 178 nett i 72 land. Televerket tilbyr også tjenestene digital og leide samband samt andre muligheter for høyhastighets dataoverføring.

Televerket har et godt utbygd mobiltelefonnett for NMT – og har i alle år vært av de ledende land i Europa. Figur 6.8 viser tetthetsgrad i mobiltelefonabonnement blant land i Europa. Med hensyn til tetthetsgrad, kvalitet og kapasitet har næringslivet i Norge, gjennom NMT, en strategisk fordel sammenlignet med de fleste land i Europa.

Televerket tilbyr også en godt utbygd personsøkertjeneste. Norge har sammen med Nederland den høyeste tetthetsgraden. Kombinert med mobiltelefon gir dette ekstra muligheter. I tillegg tilbys teksttjenesten PS-tekst.

Norge er blant de fremste land i Europa når det gjelder satellittkommunikasjon, og var det første land med satellitt i eget nasjonalt telenettet (NORSAT-A i 1976). Televerket gir næringslivet et bredt tilbud om satellittsamband, og dekker de aller fleste kommunikasjonsbehov nasjonalt og internasjonalt. Dette gjelder dataoverføring på ulike hastigheter, videokonferanse og kringkasting. Det etableres nå et VSAT-nett basert på svært små satellittjordstasjoner (antenne ned til 0,6 m i diameter).

Televerket har i dag et telenett med en infrastruktur og teknisk standard som minst er på høyde med andre nasjoner.

6.2 Organisasjon

Økt konkurranse skaper behov for en organisasjon med sterkere markeds- og kundeorientering. Målet for Televerket er å få en markedsforankret bedrift som skal betjene det norske markedet og norske kunder i internasjonal sammenheng, og som skal møte kundene der de er. Organisasjonen skal være kostnadseffektiv og ha korte og klare beslutningslinjer.

Televerket etablerer nå en ny hovedarkitektur som bygger på en konsernmodell. Etter behandling av Telemeldingen i Stortinget vil Televerket fortsatt være en forvaltningsbedrift underlagt Samferdselsdepartementet. Televerket vil eie og styre et begrenset antall datterselskaper.

Et av datterselskapene etableres som et holdingselskap, blant annet for å beskytte mot investeringsrisiko knyttet til underliggende selskaper.

Den nye hovedarkitekturen er basert på tre gjennomgående divisjoner: Nettdivisjonen, Privatmarkedsdivisjonen og Bedriftsmarkedsdivisjonen med selvstendig økonomi-, produkt- og markedsmessig ansvar. Det opprettholdes en inndeling i 7 regioner og det skilles ut egne driftsenheter i regionene.

I tillegg ønsker Televerket å etablere en egen divisjon som skal ivareta overkapasiteten i Televerket gjennom muligheter for videreutdannelse og opprettelse av nye forretningsområder.

Det er viktig at Televerket blir gitt anledning til å utvikle en organisasjon som er fullt konkurransedyktig i forhold til store internasjonale operatører.

6.3 Økonomi

Televerkets driftsinntekter i 1991 var 14 297 millioner kroner. Dette var omlag 350 milloner kroner lavere enn 1990. Solgt volum økte med 8,2 %, mens prisene på teletjenester gikk ned med 8,4 %. Resultat før ekstrordinære poster ble 1 877 millioner kroner.

Figur 6.9 viser Televerkets gjeld til Staten. Gjelden skyldes store investeringer i en årrekke før Televerket ble selvfinansiert. I 1991 betalte Televerket 2 935 millioner i avdrag til statskassen. I 1992 planlegges nedbetaling av ytterligere 3 212 millioner kroner. Målet er at Televerket ved utgangen av 1993 skal ha en gjeld på 10 milliarder kroner. Dette er fremdeles høyt sammenlignet med de øvrige nordiske land.

I 1991 investerte Televerket 3 037 millioner kroner – om lag 290 millioner kroner mer enn i 1990. Investeringene er helfinansiert gjennom avskrivningsmidler og overskudd.

Den gunstige økonomiske utviklingen har også som vist i kapittel 5, gjort det mulig å sette prisene markant ned de siste årene.

7 Internasjonalt samarbeid

7.1 Globalt samarbeid

I tråd med EFs initiativ på telekommunikasjonssektoren har også Televerket blitt lenket nærmere til det europeiske samarbeid. Arbeid innen International Telecommunication Union, ITU (CCITT og CCIR) har pågått i 125 år og fortsetter med økende effektivitet og volum (målt i antall tekstsider og rekommandasjoner!). Det samme gjør arbeidet innen andre internasjonale driftsorganisasjoner som INMARSAT, INTELSAT og EUTELSAT.

Det er imidlertid på det europeiske plan at det skjer store endringer på organisasjon og samarbeid. Televerket har i mange år vært engasjert i NORDTEL – samarbeidsorgan mellom nordiske teleadministrasjoner; likeledes i CEPT – den vest-europeiske post og telekonferanse, likeledes gjennom COST – teknisk komité for telekommunikasjoner, og i EBU – European Broadcasting Union.

Liberaliseringen og dereguleringen av det europeiske telemarkedet har nå ført til endringer.

7.2 Nordisk samarbeid

Arbeidet innen NORDTEL har de siste år stagnert. Det har blitt generert for få prosjekter samtidig som entydig forpliktelse i felles strategiarbeid savnes. Det er nå nedsatt arbeidsgrupper som skal se på formål, omfang og organisasjonsform samt vurdering av strategiske markedsforhold.

7.3 Europeisk samarbeid

CEPT ble etablert i 1959 og bestod hovedsakelig av post/teleadministrasjoner. Det pågår nå en prosess der teleoperatørdelen skilles ut av CEPT. I tillegg er teleoperatørene skilt fra teleforvaltningene. På denne bakgrunn ble forvaltningsorganet ETSI etablert i 1988. ETSI skal utarbeide telekommunikasjons-standarder innen:

- telekommunikasjon
- informasjonsteknologi (i samarbeid med CEN/CENELEC, som er europeisk standardiseringsorganisasjon)
- kringkasting (i samarbeid med EBU).

Norske medlemmer av ETSI er STF, Televerket og Alcatel a/s. De tekniske standardene fastsettes ved stemmegivninger. Standardene er i utgangspunktet frivillige, men kan ved spesielle avtaler (MOU) bli obligatoriske. Mer spesifiserte arbeidsområder er standarder for:

- telekommunikasjonsnett
- private nett, utstyr
- signaleringsprotokoller, svitsjingutstyr
- transmisjonsutstyr
- terminalutstyr
- radioutstyr og systemer
- mobilkommunikasjon, GSM
- personsøkersystem
- jordstasjoner
- avanserte testmetoder.

Innen CEPT er det også utarbeidet spesielle MOU blant annet for GSM og for ISDN.

I 1991 undertegnet 26 offentlige operatører fra 21 land en MOU gjennom grunnlegging av *ETNO,* som er et organ for å ivareta offentlige nettoperatørers interesser. Målsettingen er:

- å utvikle og representere, der det er praktisk mulig, felles holdninger i forhold til tredjepart, spesielt andre europeiske institusjoner og organisasjoner
- å samarbeide og koordinere aktiviteter mellom medlemmene for å harmonisere offentlige telekommunikasjonsnett og basistjenester i full overensstemmelse med europeiske lover.

Innen CEPT ble det i 1991 også skilt ut et eget institutt, *EURESCOM,* for forskning og strategisk planlegging. Også EURESCOM er forbeholdt offentlige nettoperatører.

Målsettingen med instituttet er å understøtte utvikling og utbygging av et harmonisert offentlig telenett og med tjenester basert på framtidsrettet forskning innen telekommunikasjonsområdet samt å utnytte dette i den europeiske integrasjonsprosess.

METRAN (Management European Transmission Network) er et CEPT-prosjekt som har til formål å realisert et digitalt bredbånds transmisjonsnett over hele Europa. Det skal bygges over synkront digitalt netthierarki (SDH) og med “cross-connect” i alle knutepunkter.

ERMES er et pan-europeisk personsøkersystem som skal realiseres i 1993. Bak systemet står 27 nettoperatører fra 18 land.

ESPRIT er et europeisk samarbeidsprosjekt fra EF-land. Det ble etablert i 1984. Formålet er:

- å fremme utviklingen av internasjonale standarder
- å skaffe IT-industrien basis teknologi
- å fremme industrielt samarbeid.

I 1985 lanserte EF-kommisjonen sitt forskningsprogram for telekommunikasjoner. Programmet fikk betegnelsen *RACE* (Research and development in Advanced Communications technologies in Europe) og har som formål å bane veien for en gradvis innføring av integrert bredbåndskommunikasjon (IBC) i Europa fra midten av 1990-årene. EF-kommisjonen vil gjennom RACE:

- styrke Europas teleindustri
- sette europeiske nettoperatører i stand til å konkurrere under best mulige betingelser
- sette flest mulig medlemsland i stand til å introdusere kommersielle bredbåndstjenester allerede fra 1995
- gi tjenesteleverandørene muligheter til å forbedre eksisterende tjenester og introdusere nye
- gjøre bredbåndstjenester tilgjengelige til konkurransedyktige priser
- støtte dannelsen av et enhetlig europeisk marked for bredbåndsutstyr og tjenester.

7.4 EFs tredje rammeprogram

Til slutt nevnes kort en oversikt over EFs tredje rammeprogram for forskning og teknologi. Totalbudsjettet for programmet er på 5.7 milliarder ECU (1 ECU = ca 8 kr). Følgende områder er dekket:

- informasjonsteknologi
- kommunikasjonsteknologi
- telematikksystemer
- industri og materialteknologi
- måling og prøving
- miljø
- havforskning
- bioteknologi
- landbruksforskning
- biomedisin og helseforskning
- biomedisinsk forskning for utviklingsland
- sikkerhet vedrørende fisjonsenergi
- kontrollert fusjonsenergi
- menneskelige ressurser og mobilitet.

8 Stortingsmelding nr 8 “Om televirksomhet i Norge”

Formålet med telemeldingen har vært:

- å gi oversikt over telesektoren i Norge
- å trekke opp utviklingslinjer framover
- å gi rammebetingelser for Televerket og Statens teleforvaltning.

8.1 Televerket fortsatt en forvaltningsbedrift

Telemeldingen er preget av den internasjonale utvikling på telesektoren og ikke minst EFs direktiver som i dette foredraget er omtalt i kapittel 3. *Telemeldingen tar imidlertid ikke de fulle konsekvenser av utviklingen* og fastholder at Televerket fremdeles skal være en statlig forvaltningsbedrift.

Situasjonen er da at teleadministrasjonene i kun tre land i Vest-Europa: Norge, Island og Østerrike, så langt forblir forvaltningsbedrifter, mens de øvrige land etablerer mer fristilte tilknytningsformer med de klare fordeler dette har i den kommende konkurransesituasjon. Østerrike vurderer for tiden annen tilknytningsform.

I telemeldingen (kap 1.1.2) er det angitt at “Televerkets økonomi og organisering må vurderes på nytt med relativt korte mellomrom”. I kapittel 4.5.2 om tilknytningsform heter det “Det bør skje en evaluering av reformene og omstillingstiltakene etter tre år. Da er det naturlig at erfaringene ..., danner grunnlag for en eventuell tilknytningsform”.

I Stortingsbehandlingen 7 april ble det imidlertid vedtatt at departementet *løpende* kan vurdere om det er behov for endringer i Televerkets tilknytningsform eller “forholdet mellom monopol og konkurransedel”.

Televerket har som forvaltningsbedrift gjennom meldingen fått utvidete fullmakter til å opprette stillinger i høye lønnstrinn, til å opprette aksjeselskaper og tegne aksjer i andre selskaper, til å belaste reguleringsfondet inntil et fastlagt minimumsnivå, og til å foreta en viss økning av investeringsutgiftene dersom dette er nødvendig.

8.2 Dereguleringsprosessen i Norge

Norge var tidlig ute med første del av dereguleringen på teleområdet. Statens teleforvaltning ble opprettet 1 juli 1987. Fra 1 januar 1988 ble det etablert fri konkurranse på terminalutstyr, blant annet telefonapparater og hussentraler. Samtidig ble TBK skilt ut som eget aksjeselskap heleid av Televerket.

Televerket har inntil nå ikke hatt normal styring med sitt datterselskap TBK slik andre bedrifter har over sine datterselskaper. Dette har hindret en mer enhetlig planlegging av tilbud på helhets telematikkløsninger for de større kunder gjennom samarbeid mellom mor- og datterselskap. Telemeldingen fastsetter nå Televerkets rett til å opprette egne aksjeselskaper på vanlig måte. Dermed vil Televerket få normal styring også over TBK. Dette ligger også inne i forslaget til den nye organiseringen av Televerket gjennom utforming av Bedriftsmarkedsdivisjonen.

I 1988 ble det også etablert fri konkurranse på verdiøkende tjenester i Norge.

Videre ble det i 1991 åpnet for full konkurranse på GSM gjennom konsesjon til TeleMobil som nå kan skilles ut i eget aksjeselskap og NetCom.

Den videre tilnærming mot framtidig konkurransesituasjon som fastslås i telemeldingen er:

- Det blir fra 1 januar 1993 tillatt med videresalg av kapasitet i leide samband.
- Det blir 1 januar 1993 full konkurranse på datex og datapak.
- Televerket forplikter imidlertid å opprettholde standardtilbud på tjenestene også i områder med liten inntjening.
- Televerket skal fortsatt ha enerett på infrastrukturen og på telefontjenesten.
- Satellitt-tjenester forutsettes å være en del av infrastrukturen og legges dermed fortsatt under forvaltningsområdet.

8.3 Tale- og dataoverføring smelter sammen

Den teknologiske utvikling fører til at tale og data kan overføres parallelt. Dette betyr at taleoverføring kan overføres over leide linjer i Televerkets egen infrastruktur. Dette er kalt bypass tidligere, men dette tillates i full konkurranse med Televerkets telefontjeneste fra primo 1993. Televerket er positive til dette vedtaket blant annet fordi det vil bidra til å løse bedrifters kommunikasjonsbehov mer effektivt, men ønsker selv å ha handlefrihet til å møte konkurransen.

Det er ingen tvil om at dette vedtaket vil få meget stor betydning for Televerkets konkurransesituasjon i tiden framover. Dette vedtaket er i full harmoni med EFs strategi for full konkurranse. Vedtaket legger forholdene til rette for at nettoperatører – små og store – kan gå inn og kjøpe opp kapasitet i Televerkets nett. Kapasiteten vil i første omgang bli tilbudt de store kundene, der den største fortjenesten ligger – altså en form for spekkhoggervirksomhet. Det er akkurat det Mercury har gjort i Storbritannia.

Televerket er nå i ferd med å planlegge og bygge ut det framtidige tjenesteintegrerte nettet, ISDN, som vil formidle tale, tekst, data og levende bilde. I kapittel 3.4 i telemeldingen heter det: "Ut fra de forslag Samferdselsdepartementet nå går inn for, skal Televerket fortsatt ha enerett på de offentlige kommunikasjonsnett og telefontjenesten, men ikke på dataoverføringstjenester. Siden ISDN-tjenester omfatter både telefon- og datatjenester, *er det etter dette lite egnet å nytte ISDN som ramme for enerett"*.

Dette betyr at telemeldingen fastslår full konkurranse på ISDN hvor telefontjenester er den viktigste. Det forventes at ISDN fra 1994 vil få en sterk vekst spesielt med tilknytninger til næringslivet og at nettet som vil være integrert med telefonnettet, blir en svært viktig del av Televerkets infrastruktur.

Det forutsettes videre at satellitt-tjenester skal være en del av infrastrukturen. Utviklingen her er igjen *i liten grad* avhengig av hva norske myndigheter beslutter, men i stedet avhengig av hvilke vedtak som gjøres internasjonalt. Dersom EF nå går inn for full liberalisering av satellittoverføringer, vil dette også i stor grad påvirke norsk telekommunikasjon – ikke minst utenlandstrafikken.

I telemeldingen oppheves nå konsesjonen på kabelfjernsyn etter telegrafloven. Det anses å skape større frihet for abonnentene med hensyn til å kunne anlegge egne kabelfjernsynsnett. Unntaket er i de tilfeller der disse nettene ønskes brukt til overføring av andre tjenester. I slike tilfeller skal det fortsatt søkes konsesjon etter telegrafloven. Departementet er ikke innstilt på å gi konsesjon til ordinære tjenester som telefon og dataoverføring, men vil utarbeide forskrifter for enkle alarmoverføringstjenester.

8.4 Fastsetting av priser på tjenester

I telemeldingen opprettholdes enhetstakster på datex og datapak uavhengig av distanse. Samtidig forpliktes Televerket til å opprettholde et geografisk standardtilbud på tjenestene over hele landet. Det skal naturligvis ikke foretas noen former for kryss-subsidiering. Rent logisk må alle forstå at slike vedtak ikke vil føre til lik konkurranse mellom Televerket og andre aktører som lettere kan komme inn og skumme fløten ved å gi tilbud til større kunder i tettbygde strøk og der distansene er korte.

Televerket har foreslått at samfunnspålagte oppgaver innen telesektoren dekkes gjennom statsbudsjettet. I telemeldingen slås det fast at Televerket med sin sterke posisjon gjennom enerett på nettet og telefontjenesten og ut fra den økonomiske utviklingen ikke trenger særskilte bevilgninger for samfunnspålagte tjenester. Dette vil imidlertid bli vurdert med jevne mellomrom.

Telemeldingen gir Samferdselsdepartementet fullmakt til å fastsette hovedtakstene på telefon slik Stortinget tidligere har gjort. I tillegg skal nå departementet også fastsette pris på leide samband – et ansvar Televerket tidligere hadde.

Televerket har altså nå mindre handlefrihet til fastsettelse av priser enn det de hadde tidligere. Som omtalt i kapittel 5 er prisstrategier og prising et svært viktig felt der Televerket må ha handlefrihet for å møte det kommende konkurransemarked.

Telemeldingen gir imidlertid Televerket anledning til å gi volumrabatter på sine tjenester.

8.5 Spesielle fullmakter

Televerket vil, innenfor rammer som foreslås i budsjettet, få utvidede fullmakter til å opprette aksjeselskap, tegne aksjer i andre selskap og avhende virksomhet, aksjer og aksjeselskap såfremt det ikke gjelder strategisk virksomhet. Strategisk virksomhet skal være heleid av staten. Det forutsettes fremdeles at både drifts- og investeringsbudsjettet er en del av statsbudsjettet. Televerket vil imidlertid få en viss frihetsgrad til å øke investeringsvolumet noe, dersom dette er nødvendig, og til å belaste reguleringsfondet inntil et fastsatt minimumsnivå.

Televerket har hatt som mål i løpet av en treårsperiode å redusere statsgjelden til 10 milliarder (ref kapittel 6). Tidligere hadde

Televerket store problemer med å få aksept for denne nedbetalingen, som var helt essensiell. Videre nedbetaling må ses i sammenheng med forventede inntekter og priser. Det er viktig at Televerket får handlefrihet til å bestemme nedbetalingstempo ut fra prisstrategier og markedsmessige vurderinger og ikke ut fra eksempelvis en saldering på statsbudsjettet.

Telemeldingen trekker fram Televerkets forsknings- og utviklingsinnsats. Det legges vekt på den betydning dette har for rask teknisk utvikling og kommersielt preget utviklingsarbeid med ringvirkninger til norsk industri og norsk IT-miljø. Det foreslås at forsknings- og utviklingsvirksomheten i Televerket økes fra 1,7 % av Televerkets driftsinntekter i 1991 til 2,0 % i 1995. Dette er i tråd med forslag fra Televerkets styre.

8.6 STFs oppgaver

Reduksjon av eneretten og større konkurranse stiller de regulerende myndigheter overfor nye oppgaver og problemstillinger. Oppgavene til Statens teleforvaltning, STF, blir i større grad å legge forholdene til rette for effektivt samvirke mellom forskjellige nett og tjenester basert på standardisering av tjenester og utstyr både i nettet og hos telebrukerne.

Det skal – gjennom Open Network Provision – sikres at teleoperatørenes infrastruktur tilbys med standardiserte grensesnitt og på ikke-diskriminerende måte. Her ligger også bestemte takstprinsipper, bruksbetingelser og radiotjenester/frekvenser.

Det europeiske standardiseringsinstituttet, ETSI, stiller strenge krav til standarder på nettinfrastruktur og viktige tjenester som vil tilbys fra ulike nettleverandører, for å sikre full samtrafikk innen EF. STF må derfor kontrollere at krav og standarder til infrastrukturen følges.

Statens teleforvaltning står også foran en mulig reduksjon i virksomheten. Det gjelder områder der gjensidig internasjonale ordninger innføres, eksempelvis på typegodkjenning. Ut fra EFTAs Tampere-konvensjon, som trådte i kraft ultimo 1990, er det innen EFTA-området åpnet for full konkurranse mellom alle typegodkjenningslaboratorier. Disse prinsippene gjennomføres nå innen EF.

Fortsatt bør STFs totale utgifter dekkes av avgifts- og gebyrinntekter. Det foreslås nå at frekvensavgiften reduseres og at det i tillegg innføres en nettoperatøravgift.

9 Virkningen av EFs innkjøpsdirektiver

Televerket kjøper materiell for store beløp – i størrelsesorden 3 milliarder kroner hvert år. Om lag 75 % av materiellet blir levert av innenlandske bedrifter. En del av verdiskapingen skjer i Norge, og Televerket må betraktes som elektronikkindustriens største kunde.

9.1 EFs offentlige innkjøpsregler

Inngåelse av EØS-avtalen vil innebære en endring i de offentlige innkjøpsregler. Dette kan være både en fordel og en ulempe for norsk industri. EFs direktiver vil åpne proteksjonistiske markeder i Europa. Dette vil gi norsk industri nye muligheter og således være en fordel. På den annen side åpnes hjemmemarkedet for mer omfattende konkurranse, hvilket kan være en ulempe.

Ifølge EØS-avtalen skal EFs direktiver innføres i perioden 1 januar 1993 – 1 januar 1995. Det er opp til norske myndigheter å velge tidspunkt for innføring. Ikke noe er bestemt foreløpig, men det er tegn som tyder på at norske myndigheter ønsker å innføre direktivene allerede i 1993 av hensyn til norsk verkstedsindustri og den mulighet dette gir for konkurranse om oppdrag på britisk sektor i Nordsjøen.

Offentlige innkjøp er definert i EFs direktiver for å fremme konkurransereglene i de statlige monopolene. Norge, og følgelig Televerket, vil bli pålagt å kjøpe til angitte kvalitetskrav for å tilpasse seg de harmoniserte grensesnitt i det europeiske telenettet. Alle televerk må forplikte seg til å legge større kjøp ut på europeisk anbud og ikke subsidiere nasjonal industri.

Reglene for offentlige innkjøp vil i sterk grad påvirke Televerket. For å sikre at direktivene etterleves av offentlige etater, har EF-kommisjonen foreslått et eget direktiv for håndheving og overvåking. Hovedprinsippene er at anskaffelse skal kunngjøres i EF, at det skal være forbud mot tekniske diskrimineringer og at objektive kriterier anvendes. CEN/CENELEC og ETSI utarbeider standarder for offentlige innkjøp. Brudd på reglene kan medføre stopp i anbudsprosedyrene, opphevelse av ulovlige vedtak og erstatning.

9.2 Terskelverdier for offentlige innkjøp

Det er angitt terskelverdier for offentlige innkjøp i ulike kategorier. Televerket ønsker *å beholde sitt gjeldende regelverk* – Regelverk for statens anskaffelsesvirksomhet, REFSA – *for innkjøp under terskelverdiene*. For innkjøp over terskelverdiene benyttes EFs direktiver.

Terskelverdiene, som ikke inkluderer merverdiavgift, er 400 000 ECU (3,2 mill kr) for varekjøp i sektorene vann- og energiforsyning samt transport. For telekommunikasjons-sektoren er det 600 000 ECU (4,8 mill kr) og for bygg- og anleggssektoren er terskelverdien 5 mill ECU (40 mill kr). Terskelverdien knytter seg til samlet verdi av prosjekt/leveranse, ikke til verdier av enkeltkontrakter på prosjekt/leveranse stykket opp i flere biter. Alle offentlige innkjøp med pris over terskelverdier, skal legges ut på anbud i databasen "Tender Electronic Daily (TED)".

9.3 Fordeler og ulemper ved EFs innkjøpsdirektiver

Fordelen til Televerket ved innføring av EFs direktiver for offentlige innkjøp ligger i markante kostnadsbesparelser ved innkjøp på grunn av større konkurranse og på grunn av de svært store innkjøp som Televerket årlig gjør.

Ulempen er at det må etableres et betydelig apparat med tilhørende systemer for å kunne håndtere det nye regelverket. Innkjøpskontoret i Televerket har anslått at et tusentalls personer i større eller mindre grad kommer i berøring med de nye EF-reglene.

Det stilles også krav til forhåndskunngjøring i Europa om kommende anskaffelser. Dette betinger 12 måneders prognoser fra ulike fagenheter, divisjoner, regioner og områder i Televerket. I tillegg må det utarbeides samleoversikter over planlagte kjøp av varer og tjenester.

Det må også foretas merarbeid med internasjonale kunngjøringer, overholding av tidsfrister, etc. Endelig kan Televerket risikere erstatningsansvar dersom direktivene ikke overholdes. Leverandører kan generelt gå til en domstol dersom de føler seg diskriminert. Dette kan i sin tur bevirke stans i en løpende anskaffelsesprosess i Televerket.

10 Hvordan møte utfordringene

Rammen rundt Televerkets virksomhet er i ferd med å endre seg radikalt. Fra å være et forsyningsmonopol skal Televerket bli en kunderettet, effektiv konkurransebedrift. Fra å ha enerett på markedene skal Televerket venne seg til å konkurrere om markedsandeler – det vil si både vinne og tape markedsandeler.

Dereguleringen innen kommunikasjonsområdet gjør at stadig flere av Televerkets enerettsområder åpnes for konkurranse. Televerket selv ønsker konkurransen, men ønsker samtidig de samme betingelser som konkurrentene.

Televerket vet at den internasjonale konkurransen etter hvert blir hardere. De store internasjonale teleselskapene står klare til å gå inn på de lønnsomme markedene.

For å møte de internasjonale teleselskapene må Televerket sikre sin egen konkurranseevne og om mulig utvikle konkurransefortrinn om forholdene tillater det.

Televerket må følge opp den internasjonale prisutvikling og foreta nødvendige prisreduksjoner for å holde på markedssegmenter.

Televerket må fortsatt være i stand til å bygge ut et moderne, sikkert og funksjonelt telenett med god kapasitet, høy framkommelighet og med et bredt tjenestespekter som er tilpasset kundenes behov.

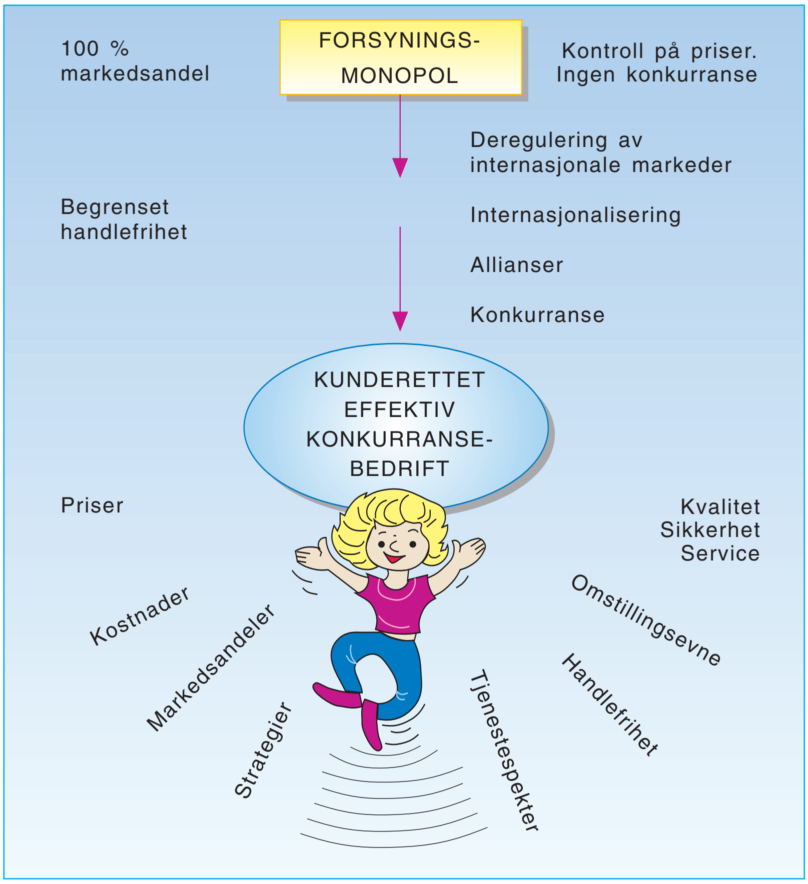
Televerket må fortsatt gi kundene høy kvalitet og service.

Televerket må foreta nødvendig produktutvikling samt være totalleverandør av teletjenester.

For å være på høyden med konkurrentene vil dette kreve at Televerket meget effektivt utnytter de ressursene som innehas.

Televerket må foreta omfattende kostnadsreduksjoner. Dette gjelder spesielt på personalsiden gjennom tilpasning til en kostnadseffektiv organisasjon. Overtallig personale skal overføres til en egen divisjon i Televerket. Driftskostnader og materiellutgifter reduseres. Nedbetaling av gjeld til statskassen fortsettes. I tillegg må det stilles strengere krav til investeringsstyring gjennom en mer profesjonell godkjenning av investeringsprosjekter.

Televerket skal øke satsingen på forskning og utvikling for å kunne utnytte resultatene i framtidige produkter og infrastruktur.



Figur 10.1 Televerkets utfordring

Televerket skal fortsatt spille en mer aktiv rolle for stimulering av norsk IT-miljø og norsk IT-industri.

Televerket skal bygge opp organisasjonen til å foreta mer internasjonal satsing for å få nye markedsandeler ute og for å følge sine kunder.

Televerket skal endre struktur til en mer markedsforankret organisasjon. Dette ivaretas nå gjennom den omfattende omorganiseringen som gjøres i Televerket.

Televerket skal endre struktur til et konsern uten topptung ledelse hvor det kan tas raske beslutninger.

Det legges vekt på at Televerket skal være en effektiv, lønnsom bedrift som er omstillingsvillig og som har sikre arbeidsplasser. Televerket skal være så konkurransedyktig, så teknologisk avansert og med så lave priser at det bidrar til å øke norsk næringslivs konkurranseevne.

I dette innlegget er det pekt på en rekke felter hvor de rammevilkår som er trukket opp av de politiske myndigheter, ikke gir den handlefrihet som Televerket snart må ha. Figur 10.1 gir en oppsummering av denne situasjonen.

En konkurransebedrift må ha mulighet til å ta raske beslutninger og ikke være i en situasjon der mange av de viktigste strategiene til bedriften behandles og diskuteres offentlig før de godkjennes. Dette er fundamentale forhold som ikke er løst ved Televerkets nåværende tilknytningsform.

Skal Televerket lykkes, må norsk næringsliv og norske myndigheter bidra til at forhold og rammevilkår på telekommunikasjonsområdet legges til rette for en sunn og likeverdig konkurranse.

Referanser

1 *Stortingsmelding nr 8, 1991-92* (Telemeldingen).

2 *Televerkets bidrag til Telemeldingen*, 1991.

3 *Televerkets kommentarer til Stortingets behandling av Telemeldingen*, 1992.

4 *Televerkets pressekonferanse om 1991 m m*, 1992.

5 *Televerkets styringsdata*, 1991-92.

6 *Televerkets årsberetning for 1992*.

7 Haga, O J. *Konkurransesituasjonen*, Televerkets plankonferanse, 1991.

8 Haga, O J. *Hva blir Televerkets rolle*. Digitalisering av telenettet, NTH, 1992.

9 Landeide, E. *Open Network Provision. Krav til tilknytningsform, grensesnitt*, 1991.

10 *Televerkets fremtid*, Teleplan, 1992.

11 Knudtzon, N. *Prospects of the Telecoms within a new European Telepolicy*, 1990.

12 Knudtzon, N. *De europeiske initiativ*. Digitalisering av telenettet, NTH, 1992.

13 *1991 EURESCOM workprogram*, EURESCOM.

14 *ETNO Informative Dossier*, ETNO, 1991.

15 EF-programmene. *Teknisk ukeblad*, 1991.

16 Hellesøe, O J. *EØS – Virkninger for Televerket av EFs innkjøpsdirektiver*, 1991.

17 Skrindo, V. *Strategianalyse for Televerket vis a vis EF 92. Innkjøpsaktivitetene,* 1991.

Etterspørsel etter 2 Mbit/s samband for telefoni i Oslo teledistrikt
– Utarbeiding av prognoser når datagrunnlaget er sparsomt

AV CARLO HJELKREM

621.39.05

1 Innledning og problemavgrensning

Store driftskostnader er ofte et problem i de eldste deler av telenettet. I Oslo teledistrikt har det siden 1990 vært lagt ned betydelig innsats for å redusere driftskostnadene i abonnentnettet.

Prognoser for abonnementsutviklingen viser at antall 30B + d aksesser vil få en kraftig vekst i perioden fram til år 2000. Slike aksesser krever 2 Mbit/s samband med tilhørende antall regeneratorer i det eksisterende nett. Tar en i betraktning at den eldste del av abonnentnettet (i sentrumsområdet av Oslo) betjener de fleste og største kundene, hvor det er flest rapporterte feil og høyest driftsutgifter, sier det seg selv at en videreutbygging med gammel teknologi er lite ønskelig.

Gruppen som har analysert dette problemet har fått aksept for ideen om å "investere seg ut av driftsutgiftene": En ny infrastruktur for lokalnettet som bygger på fiber, i stjerne eller ring avhengig av hva som er mest lønnsomt i de enkelte områder, vil på sikt gi så lave driftskostnader at slike investeringer er mer lønnsomme enn å fortsette utbyggingen med eksisterende teknologi og struktur. Dette fordi en med fiberstruktur oppnår en rekke fordeler over dagens struktur:

- Fjerning av så godt som alle *regeneratorpunkter* (i Oslo). Dette eliminerer evntuelle transmisjons-problemer i tilknytning til ISDN.
- Innføring av desentral svitsjing ved bruk av utskutte abonnenttrinn gir store besparelser av *areal, kjøling og kraft.*
- Kraftig reduksjon av antall (gamle) *kabler* i forhold til dagens trestruktur. Da får vi også færre feil, lavere vedlikeholdsutgifter og kan yte bedre service.
- Innføring av protectionsvitsjer i fiberringer gir svært god *sikkerhet:* Ved brudd svitsjes til ny overføringsvei i løpet av 5 millisekunder.
- Større *fleksibilitet og framtidig lønnsomhet* med hensyn til økning av kapasitet og oppgradering av transmisjonsutstyr.
- Bedret tilrettelegging for *bredbånds-overføringer:* B-ISDN vil direkte kunne benytte denne strukturen.
- Kunder som ønsker det kan få *fysisk atskilt reservevei* til nærmeste endesentral.
- Svitsjede tjenester blir billigere og mer konkurransedyktige overfor leide samband, fordi svitsjing og konsentrasjon av trafikken skjer lengre ned i nettet.
- Det blir lettere å *optimalisere* planleggingen av større nettområder.

Planene for utbygging av abonnentnettet er basert på et avansert planverktøy, hvor inngangsdata er *prognoser,* utstyrs- og timepriser og fysiske data fra det aktuelle området.

Utbyggingsstrategien for abonnentnettet som angis i STRATEK (8), er basert på disse prinsippene. For å kunne lage nettplaner etter disse strategiene, er det nødvendig å ha prognoser for etterspørselen etter 2 Mbit/s samband som inngangsdata.

Prognoser for antall 2 Mbit/s samband var imidlertid ikke utarbeidet tidligere og høsten 1991 meldte behovet seg for å lage prognoser for distriktets etterspørsel etter 2 Mbit/s samband i perioden fram til år 2000 (årsprognoser), fordelt på teleområder og sentralområder.

Denne artikkelen viser hvordan vi på relativt kort tid, med tilgjengelig informasjon og data, produserte et utkast til prognoser i oktober 1991.

I utgangspunktet var det ikke uproblematisk å lage slike prognoser, fordi:

- kjennskap til data og hvordan registreringssystemene behandler slike data var relativt dårlig
- den mest vanlige framgangsmåte når prognoser skal utarbeides, er å analysere historiske data. Ved hjelp av slike resultater og annen relevant informasjon, trekkes så konklusjoner om den mest sannsynlige framtidige utvikling. Eventuelle tidsrekker som beskriver den historiske utvikling vil i dette tilfellet likevel ikke være representative for den framtidige utvikling: I data-perioden har det vært beskrankninger på antall innvalgslinjer som kan tildeles (sentralenes nummerkapasitet), mens slike beskrankninger vil forsvinne med ny nummerplan.

Disse momentene bidrar, i tillegg til den usikkerhet som ligger i enhver prognose, til betydelig usikkerhet i prognosetallene.

Likevel var det, tross mulige svakheter i det endelige prognoseforslag, behov for utarbeidelse av slike prognoser: *Uten prognoser for etterspørslen etter 2 Mbit/s samband ville det ikke være mulig å gjennomføre analysen av abonnentnettet!*

Beskrivelsen av framgangsmåten er her holdt på teleområdenivå; data for hvert teleområde er benyttet i utarbeidelse av prognoser for hvert enkelt teleområde. Dette bidrar til problemer med inkonsistens når vi skal lage prognoser for hvert sentralområde etter samme metode, idet summen av sentralområdeprognosene da ikke er lik de teleområdeprognosene som vi presenterer. Dette løses vanligvis på en av to måter. Enten kan vi utarbeide sentralprognosene først og siden summere disse opp til teleområde og distrikt (bottom-up prognoser). Eller vi kan splitte allerede utarbeidede teleområde-prognoser ned på sentralområder etter en gitt nøkkel (top-down prognoser). En variant som kombinerer disse metodene er å "tvinge" summen av separat utarbeidede sentralområdeprognoser inn under en på forhånd utarbeidet teleområdeprognose. Dette skal vi imidlertid ikke gå inn på her. Det henvises til (4) hvor den siste varianten er anvendt på abonnementsprognoser. (7) behandler dette på trafikkprognoser.

Et raskt overordnet blikk på hva 2 Mbit/s samband benyttes til, viser at vi både finner slike samband på telefoni-siden, bl.a. som realisering av digitale samband for innvalgslinjer og kombinerte innvalgs/utvalgslinjer, og på data-siden (DIGITAL-tjenesten). Abonnentnettanalysen hadde således behov for prognoser for begge disse sidene.

Det vil føre for langt å gå inn på både telefoni- og datasiden her. Vi skal avgrense oss til å se på en framgangsmåte for *telefonisiden.*

Vi skal siden se at valg av data også medfører visse begrensninger på hvilke 2 Mbit/s samband vi lager prognoser for:

Prognoseforslagene som utarbeides gjelder for *"rene" digitale 2 Mbit/s samband,* dvs. hvor det er digitalt grensesnitt i begge ender av sambandet.

Dessuten gjelder prognoseforslagene kun de digitale 2 Mbit/s samband som går *mellom hussentraler og endesentraler (eller tilsvarende).* Sett fra sentralsiden, er dette tilknyttinger som krever DTM / ITM (S-12) /ETC (AXE) eller tilsvarende.

I tillegg er prognosene avgrenset til den *etterspørsel* som kommer *fra kundene,* dvs. vi har ikke inkludert Televerkets etterspørsel.

2 Datagrunnlaget

Det er tre registrerings-systemer som registrerer 2 Mbit/s samband:

1) INSA (telefoni og data)

2) S-013 (telefoni og data)

3) TELSIS

Data fra TELSIS/AB ble valgt som grunnlag for det videre arbeid. Det er flere grunner til dette, hvorav de viktigste er:

- TELSIS/AB er grunnlag for regningsutsending til abonnentene som har slike samband og anses derfor som relativt sikker mht. datakvalitet.
- Det er allerede utarbeidet prognoser for 30B + d aksesser, se (3). Disse prognosene bygger på data fra TELSIS/AB. 30B + d aksessene utgjør en viktig del av utviklingen av 2 Mbit/s samband, da slike aksesser krever slik kapasitet. Prognosene for etterspørsel etter 2 Mbit/s samband må derfor ses i sammenheng med prognosene for etterspørsel etter 30B + d aksesser.
- Kjennskap til oppbygging av abonnementsdata i TELSIS/AB er bedre enn for de andre databasene. Derfor er det naturlig å ta utgangspunkt i denne databasen.
- Utskrift fra de to andre registrerings-systemene viste relativt store avvik i registrerte 2 Mbit/s samband på teleområdenivå. Disse avvikene var det, på den tid vi hadde til rådighet, ikke mulig å finne forklaring på. Avvikene var likevel små på distriktsnivå.

2.1 Registrering av 2 Mbit/s samband i TELSIS/AB

TELSIS/AB registrerer ikke 2 Mbit/s samband direkte, bortsett fra i tilfeller med 30B + d aksesser og eventuelle MAKROTEL-abonnement, som hver krever et 2 Mbit/s samband. MAKROTEL er en relativt ny abonnementstype med samme grensesnitt som en 30B + d aksess, men uten fullt ISDN tjenestespekter ut over eget sentralområde.

Vi kan likevel komme svært nær en riktig status ved å telle opp de innvalgslinjene som TELSIS/AB har registrert, og så regne disse om til 2 Mbit/s samband.

Hvis vi tar utgangspunkt i registrerte innkoblede innvalgslinjer, finnes disse som egen post i TELSIS/AB Rapport 35: Postene Hovedabonnement (HA) og Tjenestearbeidstelefoner har "AV DISSE:" "BYLINJER TIL HUSSENTRALER/-APPARATSYSTEMER" og "INNVALGSNR.". De sistnevnte er en opptelling av antall innvalgsnummer i sentralene, og disse refererer seg direkte til innvalgslinjene mellom hussentraler og sentralen i det offentlige telenettet.

Det er i det videre arbeidet valgt å begrense analysen til de 2 Mbit/s samband som gjelder abonnementstypen HA. Dette fordi vi da får resultater som direkte er sammenliknbare med ISDN-prognosene, hvor Televerkets behov er holdt utenom.

For å analysere hva "innvalgsnr." i Rapport 35 inneholder, er det naturlig å gå til de såkalte "tariffkoder" for innvalgslinjer i TELSIS/AB. Kodene er grunnlaget for den avgift som abonnenten betaler for sine innvalgslinjer. Slår vi opp i (9), finner vi disse:

- 1031: Innvalgslinje (1 par)
- 1032: Innvalgslinje (2 par)
- 1033: Innvalgslinje (3 par)
- 1036: Innvalgslinje på PCM
- 1037: Innvalg/utvalg, ikke dynamisk.

Tariffkodene 1031, 1032 og 1033 gjelder analoge innvalgslinjer og må skilles ut fra de andre, siden 2 Mbit/s sambandene er digitale. I Oslo distrikt finnes kun abonnement på 1031 av disse tre tariffkodene.

I tillegg til disse er også to andre tariffkoder interessante, fordi de representerer abonnement/aksesser som beslaglegger 2 Mbit/s samband. Disse er ikke blant de oppførte innvalgslinjer og må derfor listes ut separat:

- 1038: MAKROTEL (innvalg/utvalg, dynamisk)
- 1076: 30B + d aksess (pilottjenesten).

Status for disse tariffkodene kan listes ut fra TELSIS/AB.

2.2 Beregning av status for antall 2 Mbit/s samband med utgangspunkt i utlisting fra TELSIS/AB

Som nevnt tidligere er det greit å oppsummere *tariffkode 1076,* idet hver 30B + d aksess krever et 2 Mbit/s samband.

Dermed gjenstår tariffkodene 1036 og 1037, hvor vi kjenner antall innvalgslinjer som kundene abonnerer på, men ikke hvor mange 2 Mbit/s samband dette utgjør.

Vi tar for oss *tariffkode 1037* først, og viser hvordan vi (tilnærmet) kan beregne antall 2 Mbit/s samband med utgangspunkt i data fra tabellene over:

For tariffkode 1037, som utgjør både inn- og utvalgslinjer, kan sammenhengen mellom antall kanaler/linjer i et 2 Mbit/s samband være til hjelp for omregning fra linjer til samband, fordi hvert 2 Mbit/s samband har kapasitet til 30 linjer. For abonnement på tariffkode 1037 kan disse således enten nyttes til linjer for innvalg eller til linjer for utvalg. En grov tilnærming til antall 2 Mbit/s samband for tariffkode 1037, ville derfor være å dividere antall linjer på hver sentral med 30. Imidlertid vil en finregning på hver kunde gi et noe høyere antall 2 Mbit/s samband, idet flere av kundene nok ikke abonnerer på alle kanalene i et slikt samband. Vi må altså gå inn på hver enkelt kunde og til slutt summere opp antall 2 Mbit/s samband som disse har. Resultatet av dette er vist i tabell 5.1.

For *tariffkode 1036* er situasjonen svært lik den som er beskrevet for tariffkode 1037 ovenfor, bortsett fra at dette kun gjelder innvalgslinjer. De eventuelle utgående linjer er altså ikke inkludert på samme måte som for utvalgslinjene under tariffkode 1037, fordi dette er analoge linjer.

Dette er i seg selv en kilde til dårlig sammenliknbare tall. Selv om analoge utgående linjer og digitale utvalgslinjer teknisk sett ikke er like, har de fra et økonomisk/nytte-synspunkt svært mye til felles. Om det er utgående linjer i kombinasjon med abonnement på tariffkode 1036 som er realisert i 2 Mbit/s samband, har vi i denne omgang ikke forsøkt å få svar på i TELSIS/AB.

Vi har altså i gjennomgangen over kun behandlet 2 Mbit/s samband som har digitale grensesnitt i begge ender av sambandet.

Våre prognoser må derfor gjelde for "rene" digitale 2 Mbit/s samband, dvs. hvor det er digitalt grensesnitt i begge ender av sambandet.

Dessuten gjelder prognosene kun de digitale 2 Mbit/s samband som går mellom hussentraler og endesentraler (eller tilsvarende), fordi det er dette data fra TELSIS/AB viser. *Sett fra sentralsiden, er dette tilknyttinger som krever DTM (S-12)/ETC (AXE) eller ITM (S-12)/tilsvarende for AXE.*

Summerer vi opp antall 2 Mbit/s samband på tariffkode 1036 på samme måte som i tilfellet med tariffkode 1037, blir resultatet som vist i tabell 5.1.

2.3 Status for antall innkoblede 2 Mbit/s samband og status for total etterspørsel etter antall 2 Mbit/s samband

I utgangspunktet er det prognoser for *etterspørsel* etter 2 Mbit/s samband som skal utarbeides. Ser vi på definisjonene som benyttes for beregning av etterspørsel etter telefoniabonnement, er total etterspørsel etter abonnement definert som antall installerte abonnement + netto antall ueffektuerte og ikke oppdaterte abonnement. Se for øvrig (4), som diskuterer definisjoner for etterspørsel. Den samme definisjonen gjelder i prinsippet for 2 Mbit/s samband.

Vi har imidlertid valgt å se bort fra denne definisjonen i denne omgang, og *antar i det følgende at antall innkoblede 2 Mbit/s samband er (og har vært) en rimelig god tilnærming til total etterspørsel.* Tre forhold bekrefter at dette er en rimelig antakelse i denne analysen:

- En rask gjennomgang av ueffektuerte og ikke oppdaterte ordrer viser et minimalt antall av slike ordrer.
- Slike ordrer gjelder pr. i dag bedriftskunder av en viss størrelse, og slike kunder blir som regel raskt behandlet.
- Analysen vil i alle tilfeller måtte gi "grove" prognoseanslag for etterspørselen.

Dessuten er det påkrevd med en grundig analyse av hva de ueffektuerte ordrene egentlig inneholder: I en del tilfeller vil dette være ordrer på utvidelser av allerede eksisterende 2 Mbit/s samband, ref. tariffkodene 1036 og 1037. En slik gjennomgang er ikke foretatt i denne analysen.

2.4 Korrigering for nummerlån

Når etterspørselen i et sentralområde skal beregnes, er det også viktig å korrigere for nummerlån, dvs. "tilbakeføre" abonnement som geografisk tilhører en sentrals dekningsområde, men som av ulike årsaker er tilkoblet annen sentral.

Slik korrigeringen gjøres ved at slike abonnement blir trukket ut av statistikken for den sentral som abonnementet er tilkoblet, og så lagt til den sentral hvis dekningsområde dekker abonnementets adresse. Dette for å få fram den reelle etterspørselen fra hvert (geografiske) sentralområde. Se (4) for en diskusjon av dette.

I dette tilfellet er slik nummerkorrigering ikke helt opplagt, fordi abonnentene i enkelte tilfeller har flyttet fra et sentralområde til et annet, men har bedt om å få beholde sine nummer i den fraflyttede sentralen. Dermed har altså abonnenten virkelig etterspurt abonnement i annet sentralområde enn det som abonnenten geografisk er hjemmehørende i.

En gjennomgang av underlaget for nummerlån mellom sentraler, fordelt på tariffkodene, viser følgende:

- 1031: 1 stk. analog innvalgslinje
- 1036: 4 stk. 2 Mbit/s samband
- 1037: 1 stk. 2 Mbit/s samband
- 1076: 5 stk. 2 Mbit/s samband.

Det var altså bare 10 stk. 2 Mbit/s samband som hadde nummerlån mellom sentraler, hvorav hovedtyngden ligger på 30B + d aksessene (1076), noe som er naturlig fordi det kun er én ISDN-sentral som betjener hele distriktet.

På bakgrunn av det som er sagt ovenfor om flyttinger mellom sentralområdene, det beskjedne antall tilfeller av nummerlån på tariffkodene 1036 og 1037 og at det i prognosen for 30B + d aksessene allerede er korrigert for dette, se (3), er det besluttet ikke å korrigere for dette.

2.5 Den historiske utvikling i datamaterialet

Innledningsvis ble det påstått at den historiske utvikling ikke vil være gyldig som grunnlag for den framtidige utvikling. Hvis denne påstanden holder, vil vi med andre ord ikke få mye hjelp fra analyse av denne utviklingen. Andre metoder og innfallsvinkler må dermed tas i bruk.

Et annet punkt er at vi heller ikke kan skaffe til veie den historiske utvikling på samme måte som vi har beregnet statusen ovenfor. Det nærmeste vi kommer, er utviklingen i antall innvalgslinjer og 30B + d aksesser fra TELSIS/AB. Omregning til antall 2 Mbit/s samband på kundenivå – slik som vist ovenfor – er ikke mulig, fordi den detaljerte historiske utvikling på dette nivå ikke tas vare på.

Historiske data vil med andre ord ikke si mye om den framtidige utvikling (når ISDN blir kommersielt tilgjengelig) og dessuten være svært grove. Dette må nødvendigvis reflekteres i valg av metode og kvaliteten på prognosene.

3 Konseptuell utvikling – antatt sammenheng mellom de ulike abonnement og aksesstypers utvikling

Før modellklasse velges, er det viktig å ha en forståelse av sammenhengen mellom de ulike linje-/abonnements- og aksesstyper. Samtidig er det viktig å se for seg den prinsippielle utvikling for disse, separat og samlet.

Utgangspunktet vårt er en antakelse om at det er 30B + d aksessene som i framtiden vil utgjøre den dominerende andel av 2 Mbit/s sambandene. De ulike tariffgruppene som er nevnt i 2.1 kan således betraktes som "generasjoner" på vei mot ISDN, dvs. 30B + d aksesser. B-ISDN er holdt utenfor vår analyse.

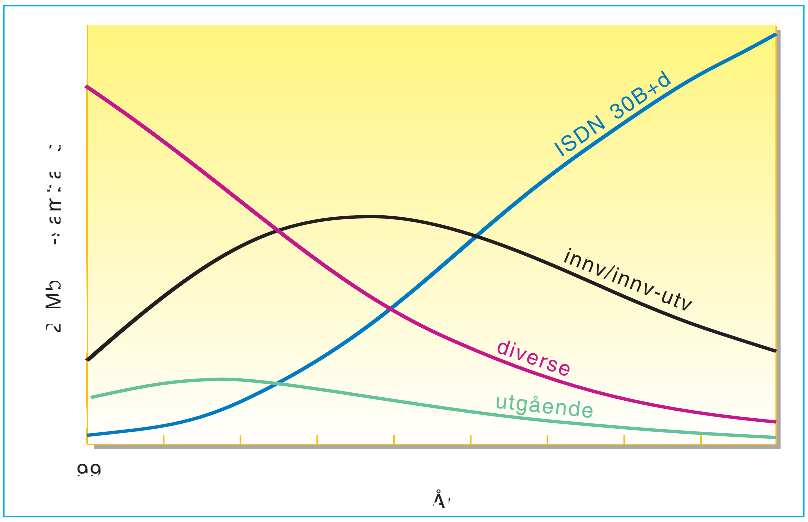
Følger vi tankegangen bak utarbeiding av ISDN-prognosene, hvor det forutsettes at abonnentene over en gitt periode konverterer sine analoge abonnement til ISDN-aksesser, kan dette lett overføres til vårt problem: *Vi forutsetter at det i prognoseperioden gradvis skjer en overgang fra eksisterende innvalgslinjer til 30B + d aksesser. I tillegg vil også inngående linjer (som ikke er innvalgslinjer) til hussentraler og enkeltabonnement konverteres.* Dette kan begrunnes ut fra lavere pris og bedret funksjonalitet i ISDN:

- *Lavere priser:* Med de prisforslag som har vært diskutert, er det stor sannsynlighet for at bruk av ISDN vil bli vesentlig rimeligere enn bruk av f.eks. dagens datanett (det samme gjelder ikke for anskaffelse).
- *Utvidet tjenestespekter:* En lang rekke tjenester vil bli tilgjengelige.
- *Dynamisk innvalg/utvalg:* Det er i dag kun MAKROTEL som gir dette. I praksis kan MAKROTEL anses som en 30B + d aksess som ikke gir fullt ISDN tjenestespekter ut over eget sentralområde.
- Begrensninger på *innvalgskapasiteten* faller bort ved innføring av ny 8-sifret nummerplan.

Alle disse forutsetningene ligger allerede til grunn for ISDN-prognosene. Den eneste forskjellen er at datagrunnlaget for 30B + d prognosene ikke er delt opp i tariffgrupper, slik som vi har gjort ovenfor.Den konseptuelle sammenhengen mellom de ulike abonnements- og aksesstypers utvikling, beskrevet ovenfor, kan illustreres som i figur 3.1.

Den blå linjen (ISDN 30B + d) utgjør ISDN-prognosen for 30B + d aksesser, slik den er utarbeidet i (3). Inkludert i denne er også MAKROTEL-abonnement, fordi slike abonnement i praksis kan anses som en 30B + d aksess. Når prognosene framkommer nedenfor, behandles disse tariffgruppene samlet.

Figur 3.1



Den sorte linjen, som i figuren er benevnt med "innv/innv-utv.", utgjør abonnement på tariffkodene 1036 og 1037. Abonnement under tariffkode 1037 er et relativt nytt tilbud, og det forutsettes at slike abonnement vil vokse i noen år, før vi får en netto overgang til 30B + d aksesser. Det er også rom for noe vekst i abonnement under tariffkode 1036. Poenget er at summen av disse to abonnementstypene forutsettes å få en utvikling som vist.

Den grønne linjen, som i figur 3.1 er benevnt med "utgående", er de utgående linjer fra abonnement under tariffkode 1036, omregnet til 2 Mbit/s samband. Dette er også abonnement som i prinsippet utgjør et potensiale for etterspørsel etter 2 Mbit/s samband: Når abonnementene på tariffkode 1036 konverteres til abonnement på tariffkode 1037, 1038 eller 1076, er det rimelig å anta at disse abonnementene også "følger med".

Selve beregningen av hvor mange 2 Mbit/s samband dette utgjør, kan gjøres ut fra følgende forutsetning: *Hver innvalgslinje på tariffkode 1036 har i gjennomsnitt 0,8 (analoge) utgående linjer.* Dette er en grov tilnærming; et sentralbord i en bedrift med forholdsvis mye inngående trafikk, f.eks. i en service-bedrift, vil ha langt flere inngående linjer (her: innvalgslinjer) enn utgående. Det omvendte vil selvsagt gjelde for et sentralbord i en bedrift med forholdsvis mye utgående trafikk, f.eks. enkelte kontor i offentlig administrasjon, som kan ha langt flere utgående linjer enn inngående. I snitt er muligens likevel denne forutsetningen ikke langt unna virkeligheten. Dette er altså en noe usikker beregning, men det er det nærmeste vi kommer sannheten med de gitte bakgrunnsdata.

Vi antar altså at noen av disse utgående (analoge) linjene først i prognoseperioden vil bli konvertert til abonnement under tariffkodene 1037 og 1038, samt at noen av linjene vil bli konvertert til 30B + d aksesser (hele perioden). Dette medfører en negativ utvikling for disse abonnementene. Hvor fort denne konverteringen vil foregå, avhenger bl.a. av abonnements/aksess-priser, avskrivningstakt på eksisterende hussentraler, samt prisutvikling på nye, kompatible hussentraler. Forutsetninger om disse forholdene er ivaretatt i prognosene for ISDN.

Den siste linjen er i figur 3.1 benevnt "Diverse". Dette utgjør en samlepott for abonnement som i dag ikke har direkte sammenheng med 2 Mbit/s samband, men som likevel antas å ville konvertere til slikt samband.

Tar vi igjen utgangspunkt i ISDN-prognosene, ligger følgende forutsetning til grunn: *Alle abonnenter som pr. i dag har 13 linjer eller flere, vil få en bedre/rimeligere løsning for sitt kommunikasjons-behov med en/flere 30B + d aksesser. For de som har færre enn 13 bylinjer/enkeltabonnement vil det, med gitt prisforslag for ISDN-aksessene, lønne seg å velge X ant. 2B + d aksesser.* Denne forutsetningen bygger på prisforholdet 1/7 for 30B + d / 2B + d, dvs. at abonnentene står overfor de samme priser ved valget mellom 7 x 2B + d aksesser eller 1 x 30B + d aksess. Dette er et av tre prisforslag som behandles i (3) og som vi velger å bruke i denne sammenhengen.

Dermed er det også av interesse å se på konvertering av abonnement ut over de som er behandlet ovenfor.

Disse abonnementene er omregnet til 2 Mbit/s samband på liknende måte som for de utgående linjene under tariffkode 1036. Underlaget for denne beregningen er ikke vist her. Vi vil her kun skissere hvordan tallene er framkommet:

Beregningen tar utgangspunkt i underlaget for ISDN-prognosene, som har talt opp alle kunder etter hvor mange abonnement de har. Status for 30B + d aksesser, omregnet til "analoge ekvivalenter", er så trukket ut av disse. Likeså er antall abonnement på tariffkodene 1037 og 1036, samt de beregnede utgående linjer fra abonnement under tariffkode 1036.

Denne "Diverse"-posten utgjør altså den resterende del av potensialet for ISDN-prognosene, og inneholder analoge innvalgslinjer (tariffkode 1031), bylinjer og enkeltabonnement. Hvor fort konverteringen vil foregå, avhenger også i dette tilfellet bl.a. av abonnements/aksess-priser, avskrivnings-takt på eksisterende hussentraler, samt prisutvikling på nye, kompatible hussentraler. Forutsetninger om disse forholdene, samt sammenhengen mellom priser og de ulike aksessvalg, er som nevnt ivaretatt i prognosene for ISDN.

Som vi ser, antar vi en avtagende utvikling for "Diverse"-sambandene: Disse konverterer til abonnement på tariffkode 1037 og 1038 i første del av prognoseperioden, mens det forutsettes en økende konvertering til 30B + d aksesser gjennom hele perioden.

4 Metodevalg

Som nevnt tidligere, vil det være naturlig å ta utgangspunkt i den allerede eksisterende prognose for ISDN-aksesser, dvs. 30B + d aksesser, når prognosene for 2 Mbit/s samband skal utarbeides. Det vil føre for langt å gå inn på detaljer om selve utarbeidelsen av ISDN-prognosene i denne artikkelen. Prinsippene for hvordan denne prognosen er utarbeidet, er det gjort rede for i (3). Vi vil ta de sist oppdaterte prognosene for 30B + d aksesser som gitt i det følgende. Dette er prognoser som er utarbeidet på et senere tidspunkt enn de som ble presentert i (3). Tallene avviker derfor noe. Utgangspunktet er som nevnt den prognosen som er basert på prisforholdet 1/7 for aksesstypene 30B + d/ 2B + d.

Det er viktig å merke seg metoden som er benyttet i utarbeidelsen av ISDN-prognosene, fordi det er klare likhetstrekk mellom datamaterialet for ISDN og for 2 Mbit/s sambandene: Vi har kun en observasjon (i ISDN-prognosene ble et scenarioanslag for observasjon nr. 2 utarbeidet) og dessuten er metningspunktet det samme, siden ISDN-aksessene en gang i framtiden vil være den eneste eksisterende tilknyttingsform, i følge forutsetningene for ISDN-prognosene.

Når vi i tillegg til resonnementet ovenfor tar i betraktning at det er "grove" prognoseanslag som skal utarbeides, er det nærliggende å velge den samme *metningsfunksjonen* som for ISDN-prognosene, som prognosemetode for 2 Mbit/s sambandene. For en enkel innføring i denne metoden, se (6). Vi velger med andre ord en modell av formen

$$Y_t = \frac{M}{(1 + e^{\alpha + \beta t})^{\gamma}} \tag{4.1}$$

hvor Y_t er etterspørsel etter 2 Mbit/s samband på tidspunkt t, M er metningspunktet, t er tiden og α, β og γ er parametre som beskriver funksjonens forløp. α-parameteren bestemmer nivået på funksjonen, mens β og γ begge bidrar til å bestemme stigningen (brattheten) på metningsfunksjonen.

Alle de fire parametrene α, β, γ og M bidrar til å bestemme formen på metnings-funksjonen. I prinsippet er det mulig å estimere alle fire, men det krever relativt avansert estimeringsverktøy. I tillegg er det en regel som sier at jo færre datapunkter (observasjoner) som er til rådighet når estimeringen skal utføres, jo færre parametre bør estimeres. Dette har sammenheng med at det er observasjonene som bidrar med den informasjon som skal til for å estimere slike parametre, og da er det lett å se at jo mindre informasjon en har, jo dårligere blir slike estimater. I dette tilfellet har vi minimalt med observasjoner: Vi har i utgangspunktet bare en, og må derfor "økonomisere" antall parametre til et minimum.

Det er relativt vanlig å estimere M utenfor modellen og så bruke denne verdien når de andre parametrene skal estimeres. Det gjør vi også her. I tillegg settes $\gamma = 20\,000$. For begrunnelse av sistnevnte, se (2). Vi står dermed tilbake med to parametre som skal estimeres: α og β. Til dette trengs minimum to observasjoner.

Når det gjelder selve metoden for estimering av parametrene α og β, er det minste kvadraters metode som benyttes her. For en innføring i denne metoden, se f.eks. (1).

Framgangsmåten blir da som følger:

1) Ta utgangspunkt i de sist oppdaterte prognoser for etterspørsel etter 30B + d aksesser.
2) Legg status for total etterspørsel etter antall 2 Mbit/s samband på tariffkode 1036 til prognosen for total etterspørsel etter 30B + d aksesser i 1991. Dette utgjør "observasjon nr. 2" i prognosen for etterspørsel etter 2 Mbit/s samband. (Abonnementene på tariffkode 1037 er allerede tatt med i 1.)
3) Startverdi (observasjon nr. 1) for 2 Mbit/s samband settes til 0.
4) Valg av startår gjøres på grunnlag av introduksjonsår for abonnementene på tariffkode 1036.
5) Metningspunktet for 30B + d aksessene beholdes, siden dette er beregnet ut fra det totale behov for 2 Mbit/s samband.
6) La modellen estimere nye verdier for α og β, og kjør prognosene fram til år 2000.
7) Ønskes prognoser som viser sammenhengen mellom prognosene i pkt. 6) og de "Utgående" linjer, (omregnet til 2 Mbit/s samband), kan status for total etterspørsel i 1991 "utvides" med disse, på tilsvarende måte som i pkt. 2 ovenfor. Pkt. 3 - 6 følges deretter som ovenfor (startår blir her det samme, siden de "Utgående" linjene er knyttet til abonnement på tariffkode 1036).
8) Ønskes prognoser som viser sammenhengen mellom prognosene i pkt. 7) og "Diverse"-posten, må denne behandles spesielt, siden startår ikke er opplagt i dette tilfellet (posten inneholder analoge innvalgs-linjer, bylinjer og enkeltabonnement). Vi har her valgt, i likhet med ISDN-prognosene, å sette et scenario-anslag for år 2000 som observasjon nr. 2. Observasjon nr. 1 blir dermed status for 1991, med dette år som startår.

2 Mbit/s	1076	1038	1037	1036	"Utgående"	"Diverse"
Oslo Sentrum Tlo	10	0	0	117	94	348
Oslo Vest Tlo	11	0	9	67	61	388
Oslo Øst Tlo	8	0	8	78	69	237
Sum Oslo distrikt	29	0	17	262	224	973

Tabell 5.1

5 Estimering av prognosemodellene. Prognoser for etterspørsel etter 2 Mbit/s samband

Vi skal i det følgende se på resultater av estimeringene, samt hvilke prognoser modellene gir.

Som en oppsummering viser tabell 5.1 status for total etterspørsel etter 2 Mbit/s samband, fordelt på de ulike tariffgrupper, slik som vist i figur 3.1. Prognosen for 30B + d (tariffkodene 1076 og 1038) i 1991 er utgangspunktet for de andre prognosene. Derfor byttes status fra TELSIS/AB ut med denne prognosen for dette året.

Tabell 5.2 viser prognosene for etterspørsel etter 30B + d aksesser og for etterspørsel etter 2 Mbit/s samband. Disse er fordelt på de tre teleområdene i Oslo teledistrikt. Vi skal i den videre gjennomgang ta for oss utarbeidelsen av 2 Mbit/s sambands-prognosene steg for steg.

Utgangspunktet er altså prognosen for etterspørsel etter 30B + d aksesser, se "1. PROGNOSE 30B + d" i tabell 5.2. Følger vi den skisserte framgangsmåten i 4, skal vi så legge status for 2 Mbit/s sambandene under tariffkode 1036 til prognosetallet for 30B + d aksesser for 1991, se tabell 5.1. Den sum som framkommer, er å finne under 1991-tallene for "2. PROGNOSE 2 Mbit/s SAMBAND: 30B + d / INNVALG / INNVALG-UTVALG".

Startverdi, null, settes til år 1986, året før introduksjon av abonnement på tariffkode 1036.

Metningspunkt for de ulike teleområdene beholdes lik de som er brukt i ISDN-prognosene.

Etter utført estimering av parametrene, gir dette følgende *prognosemodeller for total etterspørsel etter 2 Mbit/s samband i de tre teleområdene:*

Oslo Sentrum teleområde:

$$Y_t = \frac{654}{(1 + e^{1169,4 - 0,562t})^{20000}} \tag{5.1}$$

OsloVest teleområde:

$$Y_t = \frac{616}{(1 + e^{1038,4 - 0,562t})^{20000}} \tag{5.2}$$

Oslo Øst teleområde:

$$Y_t = \frac{460}{(1 + e^{116,9 - 0,566t})^{20000}} \tag{5.3}$$

Prognosene som disse modellen gir for 2 Mbit/s samband, for årene 1992-2000, finner vi under "2. PROGNOSE 2 Mbit/s SAMBAND: 30B + d / INNVALG / INNVALG-UTVALG". I sum for distriktet vokser disse altså fra 637 i 1992 til 1709 i år 2000.

6 "Avledede" prognoser

For å få en oversikt over om utviklingen for etterspørselen etter 2 Mbit/s samband, og den tilsvarende utvikling for de abonnement/linjetyper som er antatt å konvertere til slike samband, samsvarer med den konseptuelle utvikling slik den ble skissert i figur 3.1, kan vi også inkludere disse i våre beregninger. Vi følger da framgangsmåten i punktene 7) og 8) i 4, dvs. vi inkluderer status for de beregnede 2 Mbit/s sambandene under de resterende tariffkodene i status for 1991, og re-estimerer metnings-funksjonene.

For å få fram prognosetall for posten "Utgående", skal vi ifølge pkt. 7) i 4 ta utgangspunkt i status for total etterspørsel etter 2

Tabell 5.2

	Metn.pkt.	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000
OC Tlo	654	6	10	19	48	97	165	246	327	400	464	511
OV Tlo	616	7	11	21	47	99	164	238	313	382	438	483
OØ Tlo	460	5	8	15	39	75	122	178	235	285	327	359
Oslo distrikt	1731	18	29	55	134	271	451	662	875	1067	1229	1353

	Metn.pkt.	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000
OC Tlo	654	127	218	314	400	471	525	565	593	613	626
OV Tlo	616	87	161	245	327	399	457	502	535	559	577
OØ Tlo	460	94	159	227	287	336	373	400	419	432	441
Oslo distrikt	1731	308	538	785	1014	1205	1355	1466	1547	1604	1644

Mbit/s samband i 1991, se "2. PROGNOSE 2 Mbit/s SAMBAND: 30B + d / INNVALG / INNVALG-UTVALG" i tabell 5.2. Til denne status legges så tilsvarende status for 2 Mbit/s samband under "Utgående", se tabell 5.1.

Startverdi, null, settes også her til år 1986, fordi de utgående linjer følger introduksjonsår for abonnement på tariffkode 1036.

Metningspunkt for de ulike teleområdene beholdes lik de som er brukt i ISDN-prognosene.

I stedet for å presentere estimatene for parametrene innsatt i prognosemodellene, slik som i formlene 5.1 - 5.3, viser vi her kun parameterestimatene tabellarisk. Disse kan deretter settes inn i formel 4.1 slik at prognosemodellene framkommer.

Estimatene for α og β, når vi også *inkluderer "Utgående",* er gitt ved:

	Alfa	**Beta**
Oslo Sentrum tlo.:	1272,6	-0,644
Oslo Vest tlo.:	1163,4	-0,589
Oslo Øst tlo.:	1235,5	-0,651

I tabell 6.1 er statustallet for 1991, samt prognosene fram til år 2000, gjengitt under "3. PROGNOSE 2 Mbit/s SAMBAND: 30B + d / INNVALG / INNVALG-UTVALG / UTGÅENDE".

Ifølge pkt. 8) i 4, er framgangsmåten for "Diverse"-posten noe annerledes, bl.a. fordi startår ikke er helt opplagt i dette tilfellet. Dessuten er det vanskelig å anslå hva etterspørselen var i et vilkårlig valgt startår. Vi har derfor valgt å bruke status for etterspørselen i 1991 som "verdi nr. 1", med dette år som startår.

Startår settes derfor til 1991, med "verdi nr.1" i henhold til sum av 1991-tallene fra "3. PROGNOSE 2 Mbit/s SAMBAND: 30B + d / INNVALG / INNVALG-UTVALG / UTGÅENDE" og tallene fra "Diverse" i tabell 5.1.

"Verdi nr. 2", dvs. scenario-anslag for år 2000, settes til ca. 1,5 % av dagens abonnementsmasse for "Diverse"-abonnementene.

Metningspunkt for de ulike teleområdene beholdes lik de som er brukt i ISDN-prognosene.

Estimatene for α og β, når vi også *inkluderer "Diverse",* er gitt ved:

	Alfa	**Beta**
Oslo Sentrum tlo.:	487,96	-0,251
Oslo Vest tlo.:	487,75	-0,2509
Oslo Øst tlo.:	486,72	-0,2504

I tabell 6.1 er statustallet for 1991, samt prognosene fram til år 2000, gjengitt under "4. PROGNOSE 2 Mbit/s SAMBAND: 30B + d / INNVALG / INNVALG-UTVALG / UTGÅENDE / DIVERSE".

7 Oppsummering av prognoseresultater – sammenhengen mellom de ulike abonnementstypene

Prognosene er ovenfor presentert i tabell 5.2 og 6.1. For bedre å kunne vurdere resultatene, vises disse grafisk, summert for distriktet, se figur 7.1.

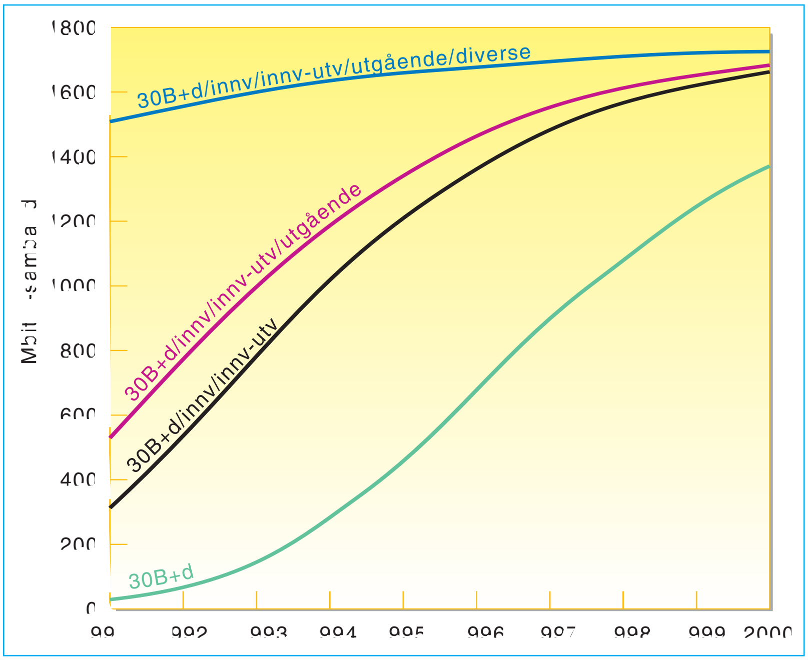
Prognosen for total etterspørsel etter 2 Mbit/s samband er sort i figuren. Nederst finner vi etterspørselsprognosen for 30B + d aksesser. Videre ser vi hvordan nivået på prognosene vokser etter som nye "andeler" av (behov for) 2 Mbit/s samband tilføyes. De to øverste kurvene representerer behov for 2 Mbit/s samband ut over de som finnes i dag. Disse har vi altså forutsatt konverteres til 2 Mbit/s samband i framtiden. I tillegg er det, gjennom anslag for potensialet for (ISDN-) prognosene, gitt rom for en viss vekst ut over det status for total etterspørsel viser i dag.

For å sammenholde prognosene med den konseptuelle skisse i 3, har vi sett på differansen mellom de ulike prognosene:

Tabell 6.1

	Metn.pkt.	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000
OC Tlo	654	209	304	390	462	518	559	589	609	624	634
OV Tlo	616	148	229	311	384	444	491	527	553	572	585
OØ Tlo	460	163	230	290	339	375	402	420	433	442	448
Oslo distrikt	1731	519	763	992	1185	1337	1452	1536	1595	1632	1667

	Metn.pkt.	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000
OC Tlo	654	569	587	601	613	622	629	634	639	642	645
OV Tlo	616	536	553	566	577	586	592	598	602	605	607
OØ Tlo	460	400	413	423	431	437	442	446	449	451	453
Oslo distrikt	1731	1505	1552	1590	1620	1644	1663	1678	1689	1699	1706



Figur 7.1

Med utgangspunkt i “4. PROGNOSE 2 Mbit/s SAMBAND: 30B + d / INNVALG / INNVALG-UTVALG / UTGÅENDE / DIVERSE” i tabell 6.1, har vi trukket fra tallene som framkommer i “3. PROGNOSE 2 Mbit/s SAMBAND: 30B + d / INNVALG / INNVALG-UTVALG / UTGÅENDE”, i samme tabell. Dermed står vi igjen med prognosen for posten “Diverse” isolert.

På samme måte har vi fra tallene i “3. PROGNOSE 2 Mbit/s SAMBAND: 30B + d / INNVALG / INNVALG-UTVALG / UTGÅENDE” i denne tabellen, trukket fra tallene fra “2. PROGNOSE 2 Mbit/s SAMBAND: 30B + d / INNVALG / INNVALG-UTVALG” i tabell 5.1. Dermed framkommer prognosen for posten “Utgående”.

Følger vi samme framgangsmåte videre, vil vi også kunne finne prognosen for “innvalg/innvalg-utvalg” isolert.

Prognosen for 30B + d aksesser var i utgangspunktet gitt, og dermed har vi nå isolert alle delprognosene. Resultatene er vist i tabell 7.1, hvor både total og årlig etterspørsel er gitt.

Tar vi for oss den årlige etterspørsel, ser vi at *“innvalg/innvalg-utvalg”* (tariffkodene 1036 og 1037), ifølge denne prognosen, får en positiv tilvekst inntil 1996. Da vil slike 2 Mbit/s samband få en negativ årlig vekst. Dette er i tråd med den konseptuelle utvikling som ble skissert i 3, idet abonnement på disse tariffkodene er forutsatt å få en del konverteringer fra postene “Utgående” og “Diverse” i første del av prognoseperioden.

Posten *“Utgående”* viser en svak positiv årlig etterspørsel i 1992. Deretter er den årlige tilveksten negativ. Den positive årlige etterspørselen i 1992 kan forklares med utvidelser av allerede eksisterende abonnement på tariffkode 1036 (det er lite trolig at vi får noen nye 2 Mbit/s samband under denne tariffkoden). Utvidelser kan selvfølgelig også forekomme i årene etter 1992, men dette vil, ifølge denne prognosen, ikke være nok til å opprettholde en positiv årlig (netto-)etterspørsel. Prognoseperioden er altså preget av konverteringer til abonnement på tariffkodene 1037, 1038 og 30B + d aksesser.

Posten “Diverse” viser en negativ årlig etterspørsel gjennom hele prognoseperioden. Dette er i tråd med den konseptuelle skissen, idet slike abonnement også konverteres til abonnement på tariffkodene 1037, 1038 og 30B + d aksesser.

8 Avsluttende merknader

De prognoseforslagene som er utarbeidet for 2 Mbit/s samband må nødvendigvis, med det bakgrunnsmaterialet som de bygger på, være grove: De gir forslag til en mulig utvikling som følger antatte prinsipielle utviklingstrekk. I praksis vil den årlige utvikling i enkelte perioder trolig kunne gi høyere etterspørsel, i andre perioder lavere. Slike nyanser vil det ikke være mulig å beskrive. Over prognoseperioden som helhet, tror vi likevel ikke at beskrivelsen av trenden i utviklingen er så ulik den vi vil få, om de gitte forutsetninger for prognosene holder.

2 Mbit/s-prognosen er svært avhengig av prognosen for 30B + d aksesser: Blir det endringer her, påvirker det prognosene umiddelbart.

Tabell 7.1

	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000
30B+d	29	55	134	271	451	662	875	1067	1229	1353
Innv/innv-utv	279	483	651	743	754	693	591	480	375	291
Utgående	211	225	206	171	132	97	70	49	33	23
Diverse	1197	1014	805	607	439	308	212	143	95	62

	1992	1993	1994	1995	1996	1997	1998	1999	2000
30B+d	26	79	137	180	211	213	192	162	124
Innv/innv-utv	204	168	92	11	-61	-101	-111	-105	-84
Utgående	13	-19	-35	-39	-35	-28	-21	-15	-11
Diverse	-183	-209	-198	-167	-131	-97	-69	-48	-33

Vi har i 2.3 (og 2.4) gjort enkelte forutsetninger og forenklinger under beregningen av status for total etterspørsel etter 2 Mbit/s samband. I tillegg er data beheftet med enkelte usikkerheter.

Svakhetene i datamaterialet kan oppsummeres slik:

1) Status for innkoblede 2 Mbit/s samband er benyttet som approksimasjon til status for total etterspørsel etter 2 Mbit/s samband. Feilen er minimal. Når ISDN "tar av" må imidlertid framtidige data vise etterspørsel, da ueffekturte og ikke oppdaterte ordrer på ISDN-aksesser trolig vil vokse betraktelig.

2) Etterspørselstallene er ikke korrigert for nummerlån mellom sentraler. Feilen antas å være liten. Framtidig arbeid må likevel vurdere og evt. bearbeide nummerlånene, da disse trolig vil vokse med kommersiell lansering av ISDN.

3) Underlagsmaterialet viser at Televerket, Oslo distrikt, står oppført med 10 linjer under tariffkode 1031. Dette er en feilkoding. Tariffkoden skulle i dette tilfellet vært 1131, og derfor ikke vært inkludert i vår liste. Slike feilregistreringer er også mulige for de andre tariffkodene.

   Andre kjente feilregistreringer er f.eks. enkelte av de første linjene med tariffkode 1036 som ble registrert. Fordi TELSIS/AB ikke var tilrettelagt for registrering av PCM-linjer, ble hver enkelt PCM-linje registrert som en linje under tariffkode 1031. Dette skal senere være konvertert til riktig tariffkode, men underlagsmaterialet viste at dette i enkelte tilfeller ikke var utført.

   Underlagsdata er ikke korrigert for slike feil.

4) Fagerborg sentral står i underlagsmaterialet oppført med 40 30B + d aksesser. Dette er feil. Under arbeidet med ISDN-prognosene ble det avslørt at TELSIS/AB i gitte tilfeller teller 30B + d aksessene feil. Det riktige antall skal være 21. Dette får imidlertid ikke påvirkning på vår analyse, siden dette allerede er korrigert i den tidligere utarbeidede prognosen for 30B + d aksesser.

5) Siden det er årsprognoser som skal lages, burde status for total etterspørsel etter 2 Mbit/s samband for årsskiftet vært lagt til grunn. Tidspunktet vi har statustall for er medio september 1991, og dermed burde et anslag for perioden ut året gjøres (evt. for perioden fra januar og fram til de avleste statustall), slik at status for et årsskifte ble utgangspunktet.

   Slik beregning er ikke utført. Under det videre arbeidet er den registrerte totaletterspørsel pr. medio september 1991 brukt som om dette var status 31.12.91. Dette medfører at tallene for de fleste av tariffgruppene nok er litt for lave som status for årsskiftet. Med henblikk på de grove prognoseanslag som skal utarbeides, er dette ansett som "rimelig", tatt i betraktning den usikkerhet som likevel ligger i analysen.

Hvis det viser seg å være behov for kontinuerlig oppfølging og revidering av disse prognosene, i likhet med abonnementsprognosene for telefoni og ISDN, bør det legges vekt på å få fram sikrere data for status, samt å følge utvikling framover.

Punktene ovenfor er områder som det ville være naturlig å "finpusse" i det framtidige prognosearbeid.

Alle disse momentene medfører at data ikke gir et helt korrekt bilde av situasjonen. Likevel vurderes underlaget som "godt nok" til et grovt førsteutkast. Videre arbeid vil måtte innbefatte en rekke forbedringer i datakvaliteten.

Takk

Takk til Johannes Bøe som har kjørt beregningene og kommet med vesentlige bidrag til artikkelen, Kjell Stordahl for nyttige kommentarer og Jan Andreas Krog som har bearbeidet datamaterialet.

Referanser

1 Bhattacharyya, G K, Johnsen, R A. *Statistical Concepts and Methods*, John Wiley & Sons Ltd., 1977.

2 Blindheim, E, Rygg M. *En modell for å utarbeide prognoser for nye teletjenester,* Rapport nr 13/82, Televerkets forskningsinstitutt.

3 Bøe, J, Hjelkrem, C. Metodebeskrivelse for utarbeidelse av prognoser for ISDN-aksesser i Oslo teledistrikt, *Telektronikk* 1, 1991.

4 Bøe, J, Hjelkrem, C. Utarbeiding av prognoser for etterspørsel etter telefonabonnement i Oslo teledistrikt, *Telektronikk* 1, 1991.

5 Log, S. *Diffusjonsmodeller for nye teletjenester*, TF-Rapport 14/82, Televerkets forskningsinstitutt.

6 Prognosekontoret, Teledirektoratet. *Bruk av metningsanalyse ved langsiktige etterspørselsprognoser etter teletjenester*, Prognosenotat 17, 1979.

7 Stordahl, K, Holden L. *Traffic forecasting models based on top down and bottom up models,* ITC 11, Kyoto, 1985.

8 STRATEK 1993. *Strategiplan for teknisk utbygging og drift,* Teledirektoratet mai 1992.

9 TELSIS-gruppen, Øst distrikt. *Bestillingstyper, tariffnummer m.m. i TELSIS,* Utgave 6, oktober 1991.

Rapport fra "15. Jahrestagung der FKTG", Berlin 1 - 5 juni 1992

AV OLAV GROV

621.397.13

FKTG (Fernseh- und Kinotechnische Gesellschaft) er en forening av firmaer og institusjoner innen produksjon, vitenskap, forskning, forvaltning og undervisning av fjernsyn og filmteknikk i Tyskland. Det er en tung forsamling av ledere og forskere med masse prestisje.

På FKTGs Jahrestagung som foregår hvert annet år er det særlig ferske forskningsresultater innen emnet som blir presentert av til dels unge representanter og grundig kommentert av forsamlingen. I år ble det lagt fram 55 foredrag.

I denne artikkelen vil jeg trekke fram en del viktige problemstillinger presentert av Frank Müller-Römer fra Bayerische Rundfunk – knyttet til framtidig digitalt jordbundet sendernett for fjernsyn – og Günther Schamel fra Heinrich-Hertz Institut für Nachrichtentechnik Berlin GmbH – knyttet til digital jordbundet koding av HDTV.

Del I: Framtidig digitalt jordbundet sendernett for fjernsyn

I begynnelsen av fjernsynstidsalderen i 50- og deler av 60-åra var det kun ett sendernett for fjernsyn i hvert land i Europa. Siden har det kommet til nye, utbyggingen av program 1, 2 og muligens 3, samt nær-TV, flere radioprogram, utbygging av kabel-TV, satellittprogram, for å nevne noen. Med utbygging av telematikknett som også vil være i stand til å overføre levende bilder, kan en spørre seg om hvor stor rolle jordbundet eterspredning av fjernsynsprogrammer vil spille i framtida.

Man går ut fra at familiens hovedmottaker i stor grad er knyttet til en fast installasjon i form av kabel-TV, satellittmottakersystem eller et telematikknett, og ekstra-apparatet i form av reise-TV fremdeles vil være avhengig av jordbundet eternett. En mener videre at den tekniske mottakskvalitet fra de jordbundne kringkastere i dag, og særlig i forhold til satellittfjernsynet, ikke lenger holder mål.

Digitalteknikkens innsats på video-området stiller kringkastingsselskaper og programprodusenter i framtida foran et grunnproblem. Er den analoge overføringen (PAL) og eventuelt (PAL-plus) konkurransedyktig i framtida? Vil brukere av jordbundet overføring havne i den situasjonen at de blir forbikjørt kvalitetsmessig av satellittfjernsyn, videorecordere, osv?

Tar man sikte på en framtidig overgang til digital distribusjonsmetode bør man både for DAB (Digital Audio Broadcasting) og DVB (Digital Video Broadcasting) stille seg følgende spørsmål og diskutere dem seriøst:

1 Antall program som skal sendes

2 Mottaksbetingelser (via telematikknett, via satellitt eller via jordbundet eterspredning)

3 Teknisk kvalitet (HDTV, TV eller noe annet).

Her har de tyske offentlige kringkastingsinstitusjonene ytret seg:

1 Beste bilde og lydkvalitet i en 7 MHz kanal for mottak med vanlig takantenne henholdsvis i en satellittkanal

2 Likekanalbruk må være mulig

3 Myk overgang (graceful degradation = langsom degradering av bildekvaliteten) ved dekningsområdets grenser

4 Det må bli tatt hensyn til videredistribusjon i kabelnett.

I forslaget til jordbundet digitalt fjernsyns-sendernett ligger det en rekke fordeler for brukerne som er av avgjørende betydning.

- Vedvarende gode mottaksmuligheter også for mobile og bærbare apparater
- Meget god teknisk overføringskvalitet
- Ingen refleksjonsforstyrrelser
- Mindre antennebehov for brukerne
- Middels programutvalg for brukerne
- Framtidssikret også for videre utvikling av fjernsynssystemer.

Systemoverveielse til Digital Video Broadcasting, DVB

Datareduksjonen blir mulig ved hjelp av spesielle metoder og algoritmer der det blir tatt hensyn til den menneskelige iakttakelsesevnen. Til tross for betydelig datareduksjon kan ikke det menneskelige øye på mottakeren se noen vesentlig forskjell fra originalen. Informasjon som vi ikke er i stand til å iaktta skal ikke overføres i datastrømmen.

Konvensjonelle kodingsmetoder for bilde, DCT = Discrete Cosinus Transformation, SBC = SubBandCoding og bevegelseskompenserende prediksjon eller en kombinasjon av disse metoder som i dag blir satt inn, har i de siste årene nådd en meget høy standard.

Hva kan vi vente oss av disse framtidige systemene? Bortsett fra Super HDTV (S-HDTV) med 4000 x 4000 bildepunkter og en datarate på ca 10 Gbit/s og med tredimensjonale muligheter, vil man i nær framtid ha to systemer klare:

System 1	625 linjer, 16 : 9 format, 7 MHz båndbredde, feilbeskyttelse og kanalkoding blir utformet og det blir mulig å motta ved en liten teleskopantenne.
System 2	1250 linjer 16 : 9 format, 7 MHz, kanalkoding for mottak med retningsantenne.

Diskusjoner om framtidige jordbundne sendernett for fjernsyn må på den andre siden ikke bli uten hensyn til hvermanns muligheter til å ta inn satellittfjernsyn direkte. I denne forbindelse kom foredragsholderen inn på betraktninger om familiens framtidige TV-utrustning. Vil f eks familiens hovedmottaker, som er en fast innretning og koblet til en fast antennekontakt, få sitt program via et jordbundet eternett? Bør HDTV og eventuelt S-HDTV distribueres utenom det jordbundne sendernettet?

Ved fastsettelse av en systemdefinisjon for digitalt jordbundet fjernsyn DVB, må en ta i betraktning følgende spørsmålsstilling og treffe tilsvarende avgjørelser:

- "Normalkvalitet" for mottak med utstyr uten tilkobling til kabel eller telematikknett (6 Mbit/s i 7 MHz kanal) og/eller

- "Low quality TV" for mottak med bærbart mobilt utstyr (3 til 5 program i en 7 MHz-kanal) og/eller
- HDTV via en utendørsantenne (30 Mbit/s i en 7 MHz kanal) og/eller
- 3D-fjernsyn via uteantenne (framtidstanker).

Digitalt jordbundet sendernett for fjernsyn

På grunn av de fysiske betingelser må man ved analogt sendernett ha stor geografisk avstand eller annen topografisk avskjerming mellom sendere på samme kanal. Bruk av digitalt likekanal-nettverk vil føre til en betydelig frekvensøkonomisering. Et program krever kun en kanal for det ønskede dekningsområdet. Det vil si at alle senderne innenfor et dekningsområde som sender samme program vil bruke samme kanal (frekvens). Hensyn til interferens vil begrense seg til andre dekningsområder som skal bruke denne kanalen til et annet program. Dette gjelder både for DAB og DVB.

Det er sannsynlig at det av hensyn til nærliggende dekningsområder (andre land) vil være mest hensiktsmessig med små og mange sendere heller enn få og store i et slikt system. Det pågår forsøk for å finne ut av dette problemet.

Man må regne med at i et digitalt nett vil det være små muligheter for mottaking utenfor dekningsområdet, slik som vi kjenner det fra vårt nåværende analoge nett. Denne muligheten må overlates til satellittkringkasting.

Videre arbeid

På samme måte som for DAB må en ha et frekvensområde å arbeide i (prøveområde). For lang tid tilbake ble kanalene 61 til 69 avsatt til slike formål. Ved en båndbredde på 7 MHz for det digitale jordbundne fjernsynsnettet betyr det for Tysklands vedkommende at kun 4 - 5 av de ni kanalene vil være tilgjengelige. Dette på grunn av nabolands frekvensbehov i samme anledning. Hvordan en kan tenke seg en senere frigjøring av disse kanaler for prøving av nye framtidige systemer vil bli avhengig av i hvilken takt en kan ta i bruk de nåværende analoge kanaler til digitalt bruk.

Avgjørende for innføringen av nye kringkastingssystemer er tilbudet av hjemmemottakere. Universalmottakere er en mulighet. Det kommer an på om en allerede i starten av forsøksperioden kan ordne seg slik at alle apparater som selges fra et bestemt tidspunkt vil være utrustet også for digital mottaking. Merkostnadene vil i forhold til fjernsynsapparatets totale kostnader være minimale.

Nasjonalt og internasjonalt blir det for tiden arbeidet intensivt på prosjekter for digitale jordbundne fjernsynssystemer:

- VADIS 625: Signalbearbeiding
- RACE T.813: Standardisering
- BMFT (HHI): Digitalt jordbundet HDTV
- Europeisk industri, teleforvaltninger, kringkastingsselskaper: Standardisering

Alle tanker om utvikling og innføring av digitalt jordbundet fjernsynssystem må ha i minne at det for tiden handler kun om rene forsknings- og utviklingsoppgaver. En innføring kan tidligst finne sted omkring årtusenskiftet. Ved alle diskusjoner om forslag og innføring av nye systemer for kringkasting må standardisering tas med i betraktningen. Her må en ta hensyn til:

- Kringkastingsinstitusjonene
- Industrien
- Abonnentene, publikum.

Bare når alle parter har fordel av innføringen av et system vil resultat kunne gjennomføres med en rimelig grad av suksess.

- Fordeler for kringkastingsselskapene?
  - · Forbedret overføringskvalitet i en 7 MHz kanal?
  - · HDTV også i et jordbundet nett?
  - · Større overføringskapasitet for flere program?
- Fordeler for europeisk industri?
  - · Masseproduksjon av de mest avanserte integrerte kretser (16 M og 64 M teknologi)
  - · 200 til 250 mill. nye fjernsynsmottakere (universalmottakere for analog/digital mottaking)
  - · 100 - 120 mill. nye videorecordere
  - · Markedspotensiale 20 til 30 milliarder DM pr år.
- Fordeler for abonnenten (publikum)?
  - · Nytt bildeformat 16 : 9?
  - · Forbedret kvalitet?
  - · Stasjonær og mobil mottaking uten forstyrrelser?
  - · Større programutvalg (der en må basere seg på mottaking via jordbundet nett).

Del II: Digital jordbundet overføring

Først generelt om jordbundet overføring:

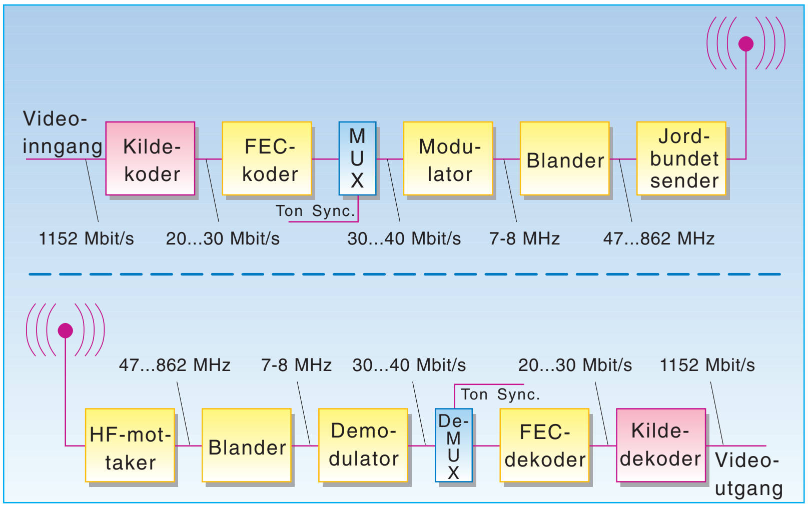
- Frekvensområdet 47 - 862 MHz, VHF/UHF kanaler med henholdsvis 7 og 8 MHz båndbredde
- 1 ... 5 bit/s pro Hz ved digital modulasjon.

Dermed får vi

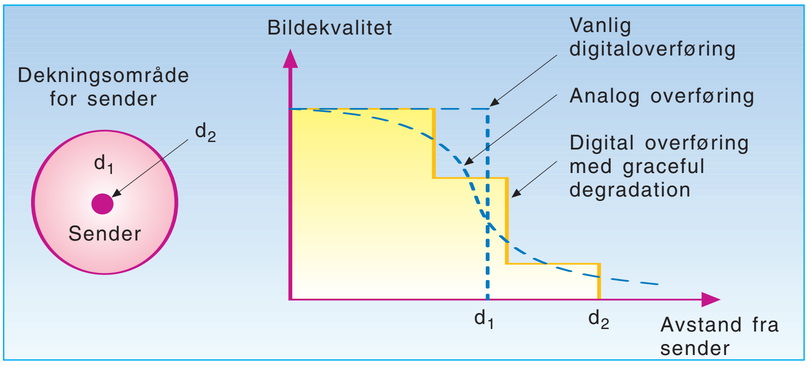
- TV med 200 Mbit/s	40 - 200 MHz
- HDTV med 1 Gbit/s	200 - 1 GHz

(I denne artikkelen opereres med tre typer TV-systemer: HDTV = High Definition TV, TV = nåværende system og PV = personal TV.)

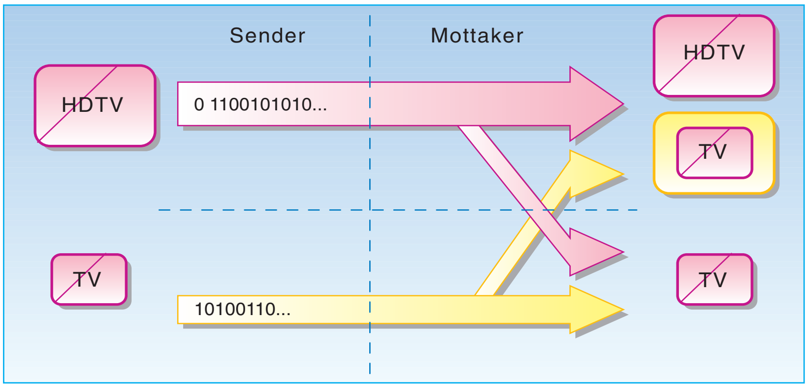
Vil man overføre en binær datastrøm over en analog kanal må det resultere i en modulasjon. For øyeblikket blir det diskutert framgangsmåter som i en Hz kan modulere datarate fra 1 til 5 bit/sek. Det betyr at vi for TV med ca 200 Mbit/s trenger mellom 40 og 200 MHz kanalbredde og for HDTV mellom 200 og 1000 MHz kanalbredde. Dermed blir det tvingende nødvendig med en betydelig datakompresjon av kildesignalet.



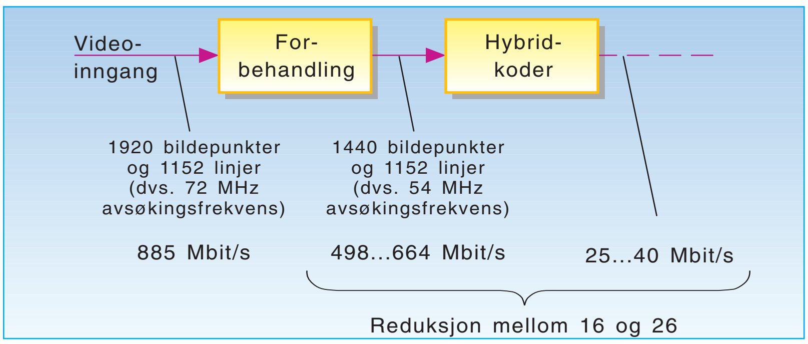
Figur 1 Overføringskjede for digital HDTV-kringkasting



Figur 2 Graceful degradation



Figur 3 Hierarkisk videokoding av HDTV-signaler



Figur 4 Forarbeiding av hybrid koding

Videosignalet blir tilført koder som reduserer dataraten til 20 - 30 Mbit/s. Videre gjennom en FEC kanalkoding. Deretter blir lyd og synk lagt til, slik at dataraten blir 30 - 40 Mbit/s. Datastrømmen blir ført gjennom en modulator som genererer et analogt signal med båndbredde 7 - 8 MHz. Dette signalet kjøres gjennom en mikser, som så gir ut den ønskede kanalfrekvensen. I mottakeren blir den inverse signalbehandlingen gjennomført.

"Graceful degradation"

Et ganske naturlig fenomen ved jordbundet overføring er at feltstyrken for mottakeren og dermed bildekvaliteten blir dårligere jo mer en fjerner seg fra senderen. Her spiller også de topografiske forhold en viktig rolle. Overfører vi over en tilsvarende sender et digitalt videosignal, får vi på mottakersiden en vedvarende god bildekvalitet så lenge den mottatte feltstyrken rekker til at den digitale datastrøm kan bli feilfritt dekodet. Ikke tilstrekkelig feltstyrke gir seg utslag i økende bitfeil og i verste fall til utfall av bildet. Dette må under alle omstendigheter bli forhindret, og det er til og med ønskelig i slike tilfeller som i den analoge overføringen å få en langsom overgang til dårligere bildekvalitet.

Det finnes tekniske løsninger innenfor FEC-kanalkodingen, men framfor alt innenfor kildekodingen. I utkastet til kodingsforløpet må dette fenomen bli tatt med.

Hierarkisk videokoding av HDTV-signaler

Et videre krav til kildekodingen er det såkalte digitale hierarki. Målet med den digitale HDTV-overføringen er å generere en bitstrøm på ca 30 til 40 Mbit/s som er bygd opp på en slik måte at en enkel mottaker, f eks TV, kan dekode deler av HDTV bitstrømmen med enkel maskinvare og likevel frambringe et brukbart bilde.

Omvendt skal HDTV-dekoderen naturligvis kunne dekode et digitalt TVs bitstrøm og framstille enten et normalt TV-bilde eller oppkonvertere til HDTV bildestørrelse.

Krav til datakompresjon

- Stor datareduksjon med god bildekvalitet (reduksjon fra 1 Gbit/s til 25 Mbit/s)
- Digitalt hierarki av videotjenester mellom HDTV, TV og PV
- Graceful degradation.

Forarbeiding av hybrid koding

Kjernen i kildekodingen består av to deler. I første trinn skjer en forarbeiding av HDTV studiosignalet. Det skjer ved en reduksjon av avsøkingsfrekvensen, slik at den opprinnelige dataraten blir redusert fra 1 Gbit/s til knappe 500 Mbit/s.

I tillegg må det i hybridkoderen foretas en reduksjon med faktor på mellom 16 - 26.

Hybridkoding av HDTV-signaler

Vi ser her et blokkskjema over en hybrid-enkoder. Her blir et beregnet bilde subtrahert fra inngangsbildet. Det beregnede bildet er framstilt av det sist dekodede bilde vha. en bevegelsesvektorkompensasjon. Så blir det beregnede feilbildet transformert, kvantisert, entropie-kodet og overført. For transformeringen blir det benyttet mest diskret cosinus transformasjon. Andre transformasjoner som f eks delbåndstransformasjon er også mulig.

Den vanskeligste delen for å realisere denne koderen er beregningen av bevegelse fra bilde til bilde.

Beregning av bevegelsesvektor

For å beregne bevegelsen fra bilde til bilde benytter man den kjente "Blockmatching"-metoden. Dette fordi det, i det minste for TV, finnes en 1-chip-løsning tilgjengelig. Derved blir for hver blokk av f eks 16 x 16 bildepunkter av et bildes n i hver blokk i bilde n - x, mest n - 1 den utvalgt som viser den største likhet. Matchingen blir dermed utført for det meste i en bestemt del av søkeområdet. Jo større søkeområde, f eks pluss/minus 32 i horisontal og vertikal retning, desto større aritmetisk mulighet for beregning av vektorene.

Prediktiv og interpolativ bildekoding

Den tidligere viste hybrid-koderen gjennomfører en prediktiv forbehandling fra bilde til bilde. Dette betyr at ut fra bare det i bildeflaten kodede bilde, et såkalt intra-bilde, blir de følgende bilder prediktive. Det vil si at de blir beregnet gjennom behandling av det siste dekodede bilde. Pilene viser derved prediksjonsretning.

Ved denne svært effektive metoden lar det seg riktignok ikke gjøre å forutsi bakgrunnen gjennom et bevegelig objekt i det aktuelle bilde, da den selv i siste bilde gjennom objektet fremdeles er skjult. Dette kan føre til kortvarig lokal opphopning av nødvendig datamengde.

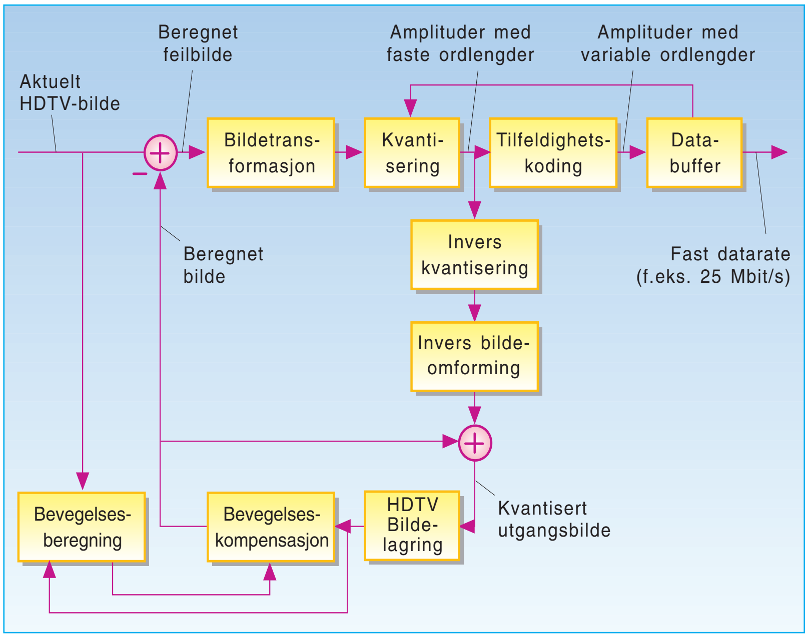
Sammenkobling av prediktiv og interpolar koding vil føre til en forbedring av dette problemet. Hva mener vi med det? Prediksjonen blir ikke overført til det neste bildet, men til et av de etterfølgende bildene, og det eller de mellomliggende bildene blir ikke tatt med.

For at en god bevegelsesoppløsning skal kunne bli regenerert i mottakeren blir de bortsatte bildene i mottakeren beregnet ved hjelp av en bevegelseskompenserende interpolasjon av de venstre og høyre nærmestliggende dekodede bilder. Da dette ikke alltid er mulig å gjøre feilfritt, blir det på sendersiden likeens foretatt en interpolasjon hvor feilen i det originale bildet som just er i senderen blir beregnet og overført. Den nødvendige dataraten er riktignok alltid betydelig lavere enn i den prediktive framstillingsmåten.

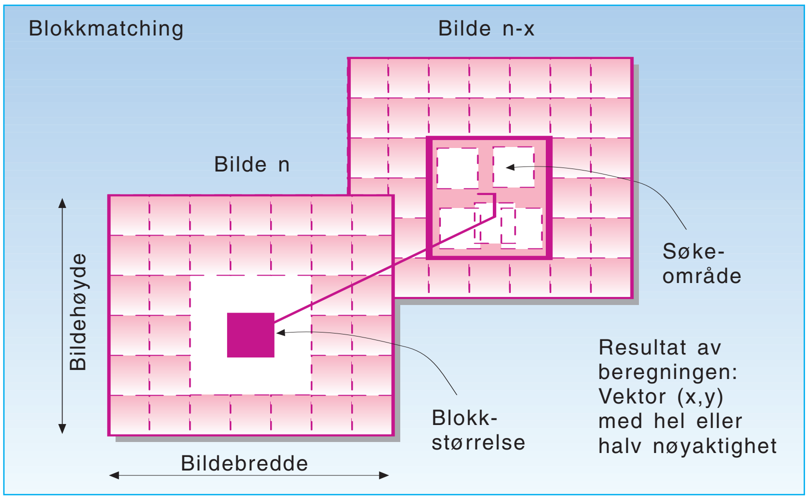
Denne metoden, som dermed samtidig er svært effektiv, har dessverre en ulempe: Mottakeren må først dekode bilde n, så bilde n + 2 før bilde n kan bli interpolert. Det betyr en ytterligere bildelagring i mottakeren og i samband med dette en utvidet totalforringing gjennom kodingen.

Realisering av det digitale hierarki

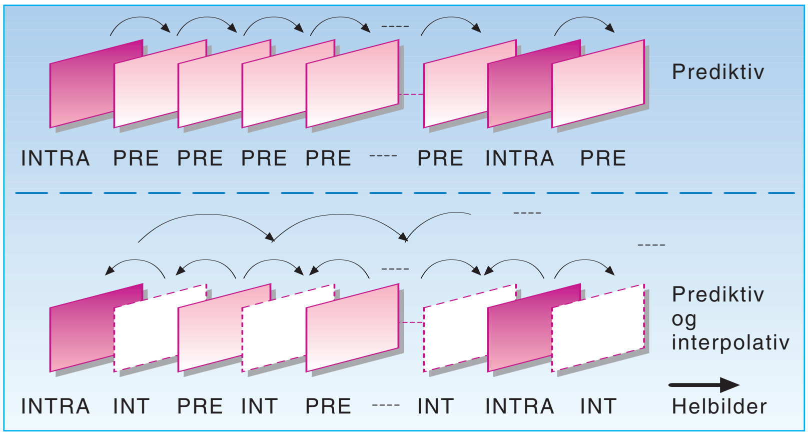
Hvordan lar det seg gjøre å nå det omtalte hierarki av kvalitetstrinn som HDTV, TV og PV, i hybrid-koderen som her er framstilt? Ser vi i tillegg koeffisientene eller den nødvendige bildetransmisjonens spektrum, kan vi bare slå fast at en del, f eks de lavfrekvente 4 x 4 DCT-koeffisientene, tilsvarer TV-andelen i HDTV-bildet. Dermed er det mulig ved 8 x 8 koeff-



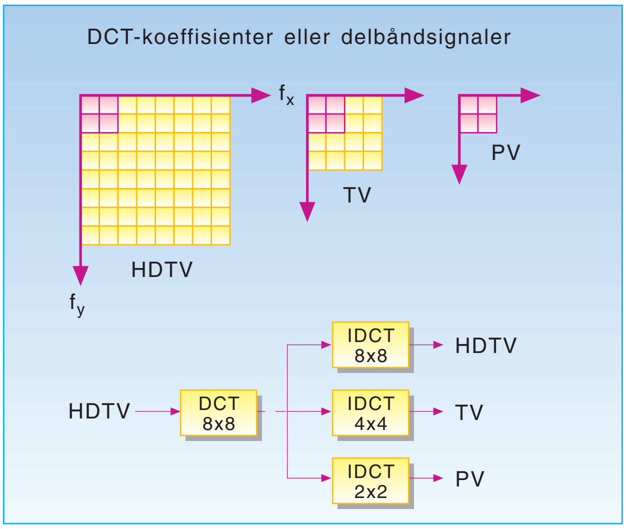
Figur 5 Hybridkoding av HDTV-signaler



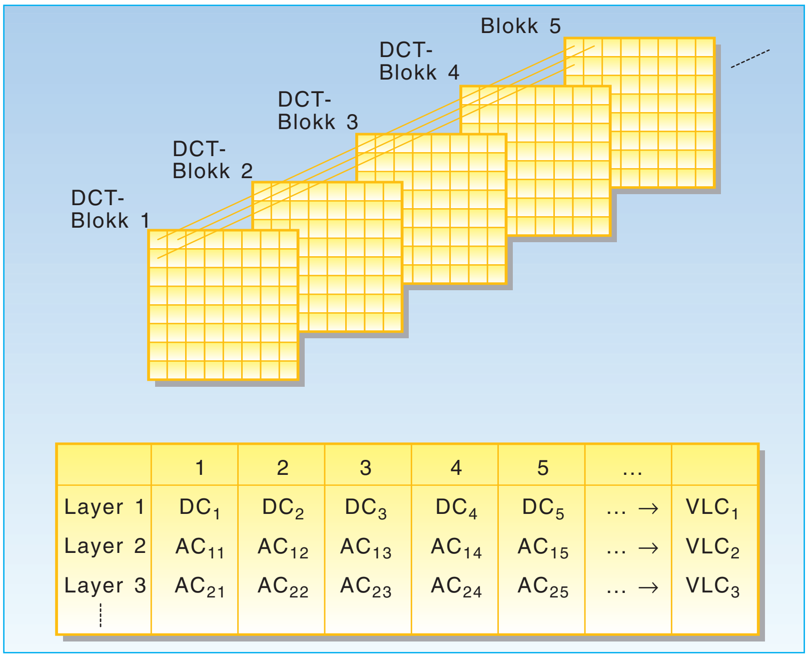
Figur 6 Beregning av bevegelsesvektor



Figur 7 Prediktiv og interpolativ bildekoding

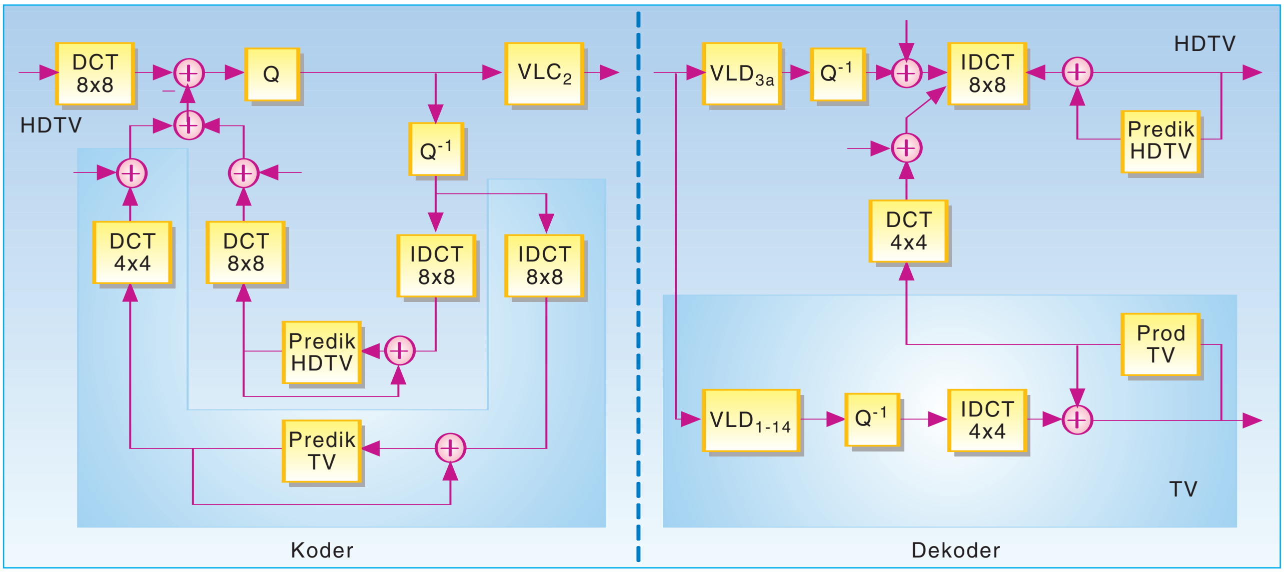


Figur 8 Realisering av det digitale hierarki



	1	2	3	4	5	...	
Layer 1	DC_1	DC_2	DC_3	DC_4	DC_5	... →	VLC_1
Layer 2	AC_{11}	AC_{12}	AC_{13}	AC_{14}	AC_{15}	... →	VLC_2
Layer 3	AC_{21}	AC_{22}	AC_{23}	AC_{24}	AC_{25}	... →	VLC_3

Figur 9 Avsøking i frekvensområdet



Figur 10 Hybrid-koding med hierarkisk prediksjon

isientene i HDTV som er beregnet i koderen å dekode alle 8 x 8, i TV-dekoderen 4 x 4 og i PV-dekoderen 2 x 2 koeffisientene.

Avsøking i frekvensområdet

En ytterligere modifikasjon må tilføres hybrid-koderen. En særlig viktig endring gjelder avsøkingen av DCT-koeffisientene som kommer etter omvandlingen og kvantiseringen. Den blir vanligvis alltid foretatt med den såkalte Zig-Zag-metoden i blokk og laget av amplitudeverdien fra den todimensjonale til en endimensjonal koeffisientrekke.

Dette er særlig uheldig for separeringen av HDTV, TV og PV. For dette, og også for kodingsforløpet, er en avsøking med like koeffisienter av de påfølgende naboblokker å foretrekke. Disse rekkene av like koeffisenter som også blir kalt "layer" (til sammen er det maksimalt 64 layer) kan med forskjellige koder lages med variable lengder. Høyfrekvente layer trenger mindre beskyttelse og kan dermed ved manglende inngangsfeltstyrke ved jordbundet overføring tåle å gå tapt uten å resultere i graverende bildefeil.

Videre må TV-dekoderen bare dekode det nødvendige layer fra 1 til 16. Dermed er både "graceful degradation" og digitalt hierarki realiserbart med den avsøkingen av frekvensområdet som her er beskrevet.

Hybrid-koding med hierarkisk prediksjon

Vi kan til nå slå fast at et digitalt hierarki ved hjelp av DCT eller "delbåndsoppløsning" er realiserbart, idet kun en del av koeffisientene blir dekodet for TVs eller PVs vedkommende. Til tross for dette vil vi med denne metoden kun få et uakseptabelt TV-bilde ved å dekode fra et HDTV-bilde.

TV-mottakeren foretar, som et resultat av sin linjefrekvens, en prediksjon som ikke overensstemmer med HDTV-prediksjonen i senderen. Det betyr at vi må modifisere prediksjonen på sendersiden slik at det tas hensyn til at det på mottakersiden også finnes TV- og PV-mottakere. Denne modifikasjonen får under ingen omstendigheter føre til en reduksjon av HDTV-bildekvaliteten i en HDTV-mottaker.

En mulighet til forbedring av TV-bildet består i å innføre en særskilt TV-prediksjon i senderen. Dette er vist i skyggeområdet på figur 10. Denne prediksjonen, som vil komme under 4 x 4 DCT koeffisienten, må i alle fall knyttes til den opprinnelige HDTV-prediksjonen på en slik måte at reduksjon av HDTV-kvaliteten ikke kan merkes når TV-kvaliteten blir forbedret.

Til høyre i figur 2 ser vi dekoder for HDTV-mottaker som også tar med TV-mottaker (skyggelagt område). TV-mottakeren dekoder kun nødvendige layer-informasjoner fra 1 til 16 som den er avhengig av.

Foredragsholderen bemerket imidlertid at dette kun er en av mange muligheter til å få fram et godt TV-bilde som for tiden blir diskutert.

Linjesprang i hybrid-koding

Går vi ut fra HDTV-halvbildene, dvs. 50 halvbilder pr sekund med hver 625 linjer pr halvbilde så får vi straks det problemet at vi får forskjellig lengde på linjerasteret i forhold til TV-signalet ved vertikal avsøking. Det finnes neppe løsninger som kan forhindre dette. Det vil i hvert fall føre til redusert kodingseffektivitet.

En enkel metode til å framstille helbilder på er å betrakte to halvbilder som et helbilde. Dette er effektivt i stillestående bilder, men i områder med bevegelser frambringer det en horisontal bevegelse fra halvbilde til halvbilde som koderen må bearbeide. Hovedfordelen er likevel den svært enkle framstillingen av TV-halvbilder av de kunstig framstilte HDTV-helbilder. Den manglende bevegelsesfase i TV lar seg kompensere med enkle midler.

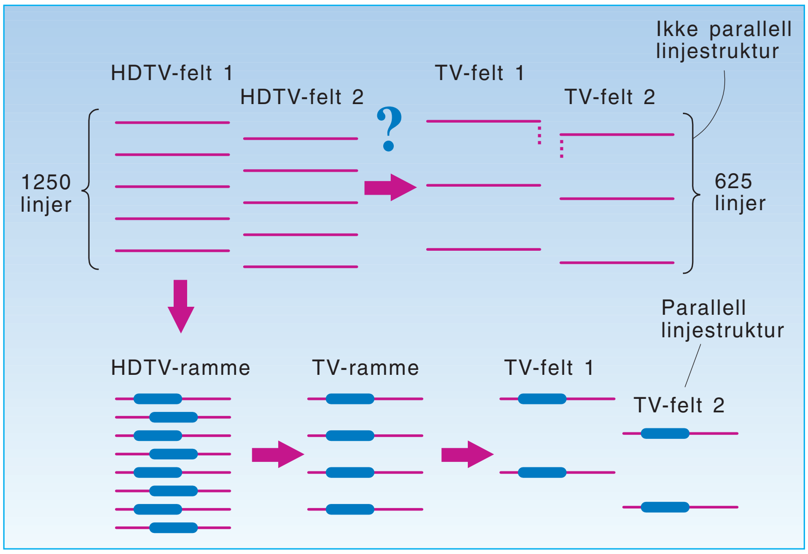
Standardisering ved MPEG/HDTV-T

La oss kort se på utviklingen av internasjonale standarder for videokoding. For tiden er vel MPEG (MPEG = Motion Picture Experts Group), en ISO-standardiseringsgruppe, den viktigste delen i dette arbeidet. For ca 4 år siden fikk de oppmerksomhet for sin videokoding av TV med 1 Mbit/s som lagringsmedium. Denne standarden er allerede blitt foreldet.

Den nåværende MPEG fase II har som mål å kode TV-tilpasset CCIR 601 med datarater mellom 4 og 9 Mbit/s. Derved forsøker man å holde seg til Generic Codings konsept. Det betyr at man må definere en algoritmes syntaks slik at de forskjellige anvendelser i de angitte datarateområdene passer inn.

De utførte simulasjonsresultater med 4 Mbit/s er forbausende gode. Bildefeil kan kun observeres i svært kritiske bildeområder. Algoritmen er på det nærmeste ferdig.

For tiden blir det mye diskutert om standarden også kan brukes for HDTV opp til 40 Mbit/s.



Figur 11 Linjesprang i hybrid-koding

Det er mulig at algoritmen kan være ferdig fastlagt i begynnelsen av 1993. Dette vil få innflytelse på overføring på jordbundet nett dersom det viser seg at algoritmen er egnet.

Det nasjonale forbundsprosjekt for HDTV-T

Til slutt ble det nasjonale forbundsprosjekt for HDTV-T presentert, noe som Heinrich-Hertz Institut har vært med på å blåse nytt liv i. Dette prosjektet har satt seg som mål å realisere et digitalt jordbundet overføringssystem som skal være klart for feltforsøk i løpet av 1995.

Oppsummering

Sammenfatningsvis kan en formulere kravene til datakompresjon for digitalt overføringssystem som følger:

1 Svært god bildekvalitet

2 Graceful degradation

3 Digitalt hierarki ved hjelp av en skalerbar datastrøm.

Standard for videokoding

- Fastsettelse i begynnelsen av 1993
- Egnet for jordbundet overføring?

Internasjonalt er MPEG krumtappen. Hele verdens dataindustri og kringkastingsselskaper følger nøye med i framskrittene. Det er mulig at standarden kan brukes til digital jordbundet overføring.

Referanser

Müller-Römer, F. *Künftige digitale terrestrische Sendernetze für Fernsehen.*

Schamel, G. *Codierung für die digitale terrestrische HDTV-Übertragung.*

15. Jahrestagung der FKTG, Berlin, 1 - 5 juni 1992.

Tale kontra data og video

AV INGE VABØ

621.39.05

I St.prp. nr 1 (1992-1993) er Televerkets budsjett for 1993 lagt fram for Stortinget. Investeringsbudsjettet for 1993 er redusert med 11,1 % i forhold til 1992 og forslaget lyder på 3,19 milliarder kroner. Investeringsmidlene må Televerket selv finansiere. Investeringene skal gi Televerket inntekter i framtiden. Inntektene vil primært måtte komme fra trafikken som kundene genererer. I St.prp. nr 1 er volumveksten for innenlandstrafikken forventet å stige med ca 5 % pr år og utenlandstrafikken med 8 % pr år fram til 1996. Det forventes en gjennomsnittlig økning av antall hovedabonnenter på ca 60 000 pr år (under 3 %). I tillegg vil antall mobiltelefonabonnenter øke, men det er ikke anslått noe tall for dette (usikkert pga. konkurransen). I tillegg til telefontrafikken vil data og video generere økning av trafikken i årene framover. Denne trafikken tror Televerket vil øke mye raskere enn telefontrafikken og planlegger derfor et bredbåndsnett for å ta denne. Men hva er det man snakker om av kapasitet i forhold til telefontrafikken?

Vi skal se litt på dette:

Telefontrafikk

Det er 2,2 millioner abonnenter i Norge i 1992. Hvis vi antar at det ringes 10 samtaler pr abonnent pr dag (flere brukere pr telefon) med 3 minutters varighet pr samtale, blir dette $2{,}2 \cdot 10^6 \cdot 3 \cdot 10 = 66$ millioner minutter/dag. Hver samtale krever 64 kbit/s · 2 (veis) = 128 kbit/s kapasitet.

Dette skulle da gi pr år:
$66 \cdot 10^6 \cdot 128 \cdot 10^3 \cdot 60 \cdot 365 = 185 \cdot 10^{15}$ bit.
(10^{15} = 1000 billioner)

Vi kan nå sammenlikne denne telefontrafikken med data- og videooverføring:

Flyreservasjoner

Det fløy totalt 16 millioner passasjerer innenlands og utenlands fra Norge i 1990. Terminaler benyttes til reservasjonene. Anta at det kreves 30 bildeskjermer informasjon for hver reservasjon. Hvert skjermbilde inneholder 2500 tegn à 8 bit/tegn.

Dette gir da:
$16 \cdot 10^6 \cdot 30 \cdot 2500 \cdot 8 = 0{,}01 \cdot 10^{15}$ bit.

Dette var sørgelig lite sammenliknet med telefontrafikken.

Informasjonsarbeidere

Det er ca 2 millioner arbeidstakere i Norge. Vi antar at i framtiden blir alle informasjonsarbeidere som sitter foran sine terminaler hele dagen. La oss si at de klarer å "fordøye" 200 skjermbilder pr time og hvert bilde inneholder 1000 tegn (bytes) informasjon (1 byte = 8 bit).
Hvis de arbeider 8 timer pr dag 200 dager i året blir dette:
$2 \cdot 10^6 \cdot 200 \cdot 1000 \cdot 8 \cdot 8 \cdot 200 = 5 \cdot 10^{15}$ bit.

Dette er også lite sammenliknet med telefontrafikken.

Persondata

Det bor 4,2 millioner mennesker i Norge. La oss anta at av alle data samlet om oss (bankdata, forsikringsdata, folkeregisteret, osv), blir det sendt over telenettet 10 000 bytes om dagen.

Dette gir:
$4{,}2 \cdot 10^6 \cdot 10\,000 \cdot 8 \cdot 365 = 0{,}12 \cdot 10^{15}$ bit.

Telefontrafikken vinner suverent!

Levende bilder

Dette må vel ta innersvingen på telefontrafikken? La oss anta at vi har 10 TV-kanaler som skal overføres over telenettet 8 timer pr dag. Uten noe særlig koding vil hver TV-kanal kreve 100 Mbit/s.

Dette gir:
$10 \cdot 100 \cdot 10^6 \cdot 8 \cdot 3600 \cdot 365 = 10{,}5 \cdot 10^{15}$ bit.

Dette er bare vel 5 % av telefontrafikken!

Aviser

Vi kan tenke oss at avisene i Norge kan bli distribuert gjennom telenettet. Det er 150 aviser tilsluttet Avisenes Landsforening. Av disse kommer ca 60 ut daglig og 5 kommer også ut om søndagen. Det er ca 50 aviser i tillegg som kommer ut én eller to ganger i uken. La oss derfor anta 100 aviser pr dag. La oss anta at hver avis har gjennomsnittlig 20 sider. Skal også bildene i avisen overføres med kvalitet må det overføres minst 2 Mbit/s pr side (kodet).

Dette gir:
$100 \cdot 20 \cdot 2 \cdot 10^6 \cdot 365 = 0{,}001 \cdot 10^{15}$ bit.

Selv om vi antar at det må retransmitteres pga. feil og korreksjoner, blir det et forsvinnende antall bit sammenliknet med telefontrafikken.

Blader

Blader og tidsskrifter har mer bildestoff og krever derfor 12 Mbit/s pr side. De inneholder også flere sider – ca 100 i gjennomsnitt. Ifølge Bladsentralen utgis ca 30 pr uke.

Dette skulle gi:
$30 \cdot 12 \cdot 10^6 \cdot 100 \cdot 52 = 0{,}02 \cdot 10^{15}$ bit.

Dette er heller ikke mye å skryte av!

Datamaskinoverføringer

La oss anta at det er 600 000 PC-er i Norge. La oss videre anta at disse hver skal ha overført 10 Mbyte med informasjon pr dag.

Dette gir:
$600\,000 \cdot 10 \cdot 10^6 \cdot 8 \cdot 365 = 18 \cdot 10^{15}$ bit.

Dette er fremdeles bare knapt 10 % av telefontrafikken.

Videotelefon

La oss anta at 10 % av telefonabonnentene skaffer seg videotelefon. Da må takstene bli lave! Anta at videobildet er

komprimert til 64 kbit/s. Hvis de ringer like mye som den vanlige telefonabonnent, 10 ganger pr dag med 3 minutter varighet pr gang, blir dette:
$220\,000\ (10\ \%) \cdot 10 \cdot 3 \cdot 64 \cdot 10^3 \cdot 2 \cdot 60 \cdot 365 = 18{,}5 \cdot 10^{15}$ bit.

Det er ikke sikkert at abonnentene vil like å se hverandre hver gang de ringer, så behovet for kapasitet er optimistisk!

Sluttbemerkninger

Ringer vi i 3 minutter genererer vi 11,5 Mbit med informasjon over en 64 kbit/s linje. Dette tilsvarer 1 400 000 tegn!

Hvis vi summerer alle data- og videoinformasjoner ovenfor får vi:
$0{,}01 + 5 + 0{,}12 + 10{,}5 + 0{,}001 + 0{,}02 + 18 + 18{,}5 = 52{,}15 \cdot 10^{15}$ bit.
Dette er bare ca 1/3 av telefontrafikken!

Nesten 80 % av telefontrafikken er lokal trafikk. Fordeling av framtidig data- og videotrafikk mellom lokal- og fjerntrafikk er ukjent.

Datatrafikken sendes som kjent i "bursts" (skurer). Kapasitetsbehovet kan være stort i korte perioder. Dette vil forårsake at Televerket må bygge ut "overkapasitet" som går på tomgang mesteparten av tiden. Dette er svært kostbart og er det nødvendig? Både data- og telefontrafikk kan ved hjelp av kompresjonsteknikk redusere kapasitetsbehovet. Telefonsamtaler kan i hvert fall overføres med 32 kbit/s uten at man merker kvalitetsdegradasjon.

Det synes som om det er telefontrafikken som blir den helt dominerende inntekten for televerkene også i framtiden. Derfor må transmisjonsnettene dimensjoneres etter veksten i telefontrafikken, og vi bør ikke bli lurt til overinvesteringer basert på video- og dataprognoser som vil gi et nett som går på "tomgang" mesteparten av tiden.

Vi må ved hjelp av takstpolitikken fordele trafikken over 24 timer i døgnet slik at investeringene i nettet kan bli bedre utnyttet.

B-ISDN - et viktig grunnlag for multimedia

AV KJELL HERMANSEN

621.39.05

Resymé

Artikkelen er en gjennomgang av muligheter som B-ISDN gir for overføring av multimedia informasjon.

1 Introduksjon

Ordet medium betegner ulike ting i ulike sammenhenger. Vi skal her bruke det for å betegne en informasjonskanal med en bestemt *presentasjonsform.* Mennesker kan f eks presenteres informasjon vha grafiske tegn, bilder, tale – maskiner vha signaler. Disse informasjonskanalene sier vi her er ulike media.

Teleterminaler har til nå gjerne vært spesialtilpasset ett medium (f eks er telefon tilpasset tale, telex er tilpasset tegn og telefax tilpasset bilder (f eks bildet av en side med tekst)). Telenettene har gjerne også vært tilpasset de krav som disse terminaler satte. Slik kan en si at offentlige nett og terminaler i stor grad har vært fokusert mot spesielle media. Brukerne har gjerne via sine terminaler fått tilgang til *monomedia* tjenester.

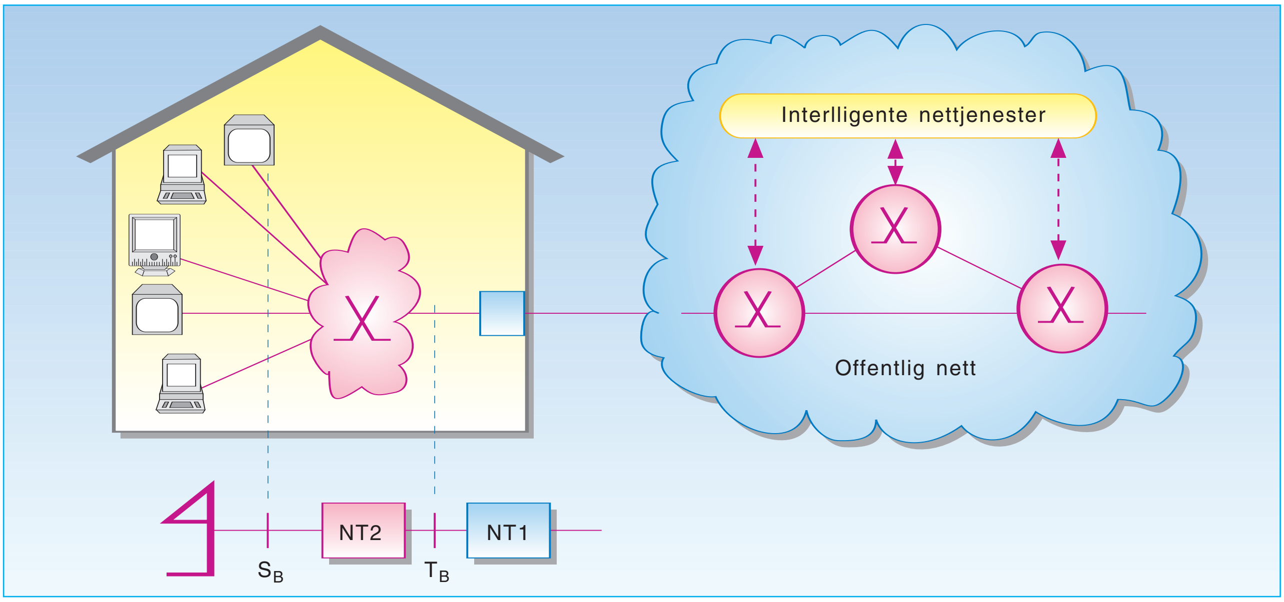
Utviklingen innen informasjonsteknologi gir økt tilgang til systemer hvor flere media benyttes samtidig (f eks bildetelefon og tilgang til superdatabaser (databaser hvor informasjonen er lagret som en kombinasjon av tale, video og tegn)). Det blir derfor stadig viktigere at det enkelte nett kan overføre informasjon fra flere media: Teleadministrasjonene oppnår rasjonaliseringsgevinst siden de ikke behøver å operere med flere parallelle nett – og det er praktisk for brukerne å kunne benytte en og samme nett-tilknytning når flere media benyttes.

Telefonnettet var opprinnelig et spesialnett for taleoverføring, men etter hvert har det også blitt tatt i bruk til formidling av andre media. De begrensede overføringsegenskapene til telefonnettet er forbedret i ISDN. ISDN er et kortord for Integrated Services Digital Network, og ISDN er nettopp utviklet med tanke på at brukere skal ha tilgang til flere enkelt-media tjenester via samme nett-tilknytning – gjerne også fra samme terminal. Brukeren sies da å ha tilgang til en *multimedia* tjeneste.

Det er fundamentalt for moderne telekommunikasjon at informasjon knyttet til ulike media kan kodes som sekvenser av bit (digitaliseres). Ulike media kan sette svært forskjellige krav til nettets evne til å overføre bit. (Kravene omfatter: evne til overføring av et gitt antall bit innen et tidsrom, maksimal toleranse når det gjelder variasjoner i overføringshastigheten og feil i overføring av bitsekvensene.) Bit-overføringskravene for det enkelte medium vil også variere sterkt avhengig av hvor mye av informasjon (f eks i et bilde) som overføres og hvordan den kodes.

Hvorvidt kravene fra de ulike media skal bli synlige i nettet, er sterkt avhengig av tiden en har til rådighet fra en begynner å overføre et informasjonselement (f eks et tegn) til elementet skal *presenteres* hos mottakeren. Dette kalles sanntidskrav. Ved elektronisk post kan hele meldingen (f eks inkludert tale) være overført gjennom det offentlige nettet før brukeren får den presentert. Dersom sanntidskravet – slik som i dette tilfellet – er svakt, kan overføringstiden være lang, og da kan også kravene til nettets andre overføringsegenskaper svekkes. (Kravet til nøyaktighet kan f eks reduseres fordi feilaktige bitsekvenser kan sendes om igjen.) Hovedsaken blir da at informasjonsoverføringen totalt sett kan gjøres billig for brukeren. Dess sterkere sanntidskrav – dess mer må kravene fra de enkelte media også tas i betraktning ved bitoverføringen gjennom nettet.

B-ISDN er nettkonsept beskrevet i internasjonalt godkjente forskrifter (anbefalinger fra CCITT) for overføring av digitalisert informasjon gjennom offentlige nett. Vi skal i denne artikkelen drøfte hvilke muligheter B-ISDN generelt gir for overføring av bitstrømmer fra ulike media – både når sanntidskravene er sterke og når de er svake.



Figur 1 B-ISDN nettstruktur

2 Nettkonsept for multimedia som omfatter offentlig og privat utstyr

B-ISDN anbefalingene omfatter både regler for utstyr som tilknyttes direkte til det offentlige nettet og utstyr som tilknyttes via spredenett ute hos abonnentene. B-ISDN beskriver også den interne nettoppbyggingen.

Figur 1 viser viktige elementer som inngår i B-ISDN – og en mer formell framstilling (referansemodell) for nett-tilknytningen.

Via standardiserte grensesnitt (S-grensesnitt) får terminaler tilgang til B-ISDN nett-tjeneste. Spredenettet (NT2) ute hos brukeren kobles også til det offentlige nett via et standardisert grensesnitt (T-grensesnitt). Det offentlige nett avsluttes med et utstyr (NT1) ute hos abonnenten som overvåker at informasjonsoverføringen skjer korrekt – og eventuelt omformer lavere lags protokoller dersom protokollene på linja inn til abonnenten er forskjellig fra det standardiserte T-grensesnitt.

Både på de nevnte grensesnittene og internt i nettet benyttes Asynchronous Transfer Mode, ATM, som overføringsmetode. Viktige prinsipper er her at informasjonen overføres vha bitsekvenser (celler) som alle har samme lengde, og at cellene også inneholder informasjon om hvilken kanal som brukerinformasjonen tilhører. B-ISDN gir slik mulighet til å opprette flere informasjonskanaler over samme fysiske linje.

En er for tiden opptatt med å definere S-grensesnittet i B-ISDN, slik at flere terminaler kan tilknyttes fra samme uttak i NT2. (Den enkelte terminallinje blir da å oppfatte som fellesmediet i et LAN, Local Area Network.) Men det reises også spørsmål om fellesmedium-løsninger vil være kosteffektive ved informasjonsoverføring med sterke sanntidskrav.

I tillegg til rene overføringstjenester vil brukeren i B-ISDN også få tilgang til intelligente nett-tjenester som f eks medflytting av abonnement.

3 Med B-ISDN vil investeringer i fiber kunne utnyttes til å møte utvidede overføringskrav fra multimedia

Narrowband ISDN, N-ISDN, er første generasjon av ISDN og gir abonnentene mulighet til å sette opp flere digitale forbindelser over samme grensesnitt. Disse forbindelsene er begrenset til 64 kbit/s, og denne kapasiteten beslaglegges i det utstyr i nettet som forbindelsen etableres gjennom. Abonnenten har også tilgang til en signaleringskanal som kan benyttes til å kommunisere med nettet (f eks for å få satt opp forbindelser). Slik får abonnenten tilgang til flere informasjons*kanaler* han kan benytte samtidig.

Brukeren kan kommunisere med flere brukere samtidig gjennom kanalene ved å koble dem opp til spesielle noder (servere) og benytte en pakkeorientert nettprotokoll (f eks forbindelsesfri nettprotokoll) mot disse.

Den kapasitet som tilbys vanlige brukere er relativt begrenset: to 64 kbit/s kanaler og en 16 kbit/s signaleringskanal. (Ledig kapasitet i signaleringskanalen kan også benyttes til datatrafikk via en server.) Allerede i dag ser en at applikasjoner som krever levende bilder med god kvalitet – og høyhastighets datakommunikasjon, vil kreve høyere overføringskapasitet enn den som N-ISDN kan gi.

Neste generasjon av ISDN, B-ISDN, tar sikte på å gi abonnenten mulighet til å benytte flere kanaler (16 777 216) og gir ham større overføringskapasitet. Abonnenten kan enten få tilgang til 135 631 kbit/s eller 542 526 kbit/s til sin informasjonsoverføring. (Asymmetrisk nett-tilkobling med 135 Mbit/s i den ene retning og 542 Mbit/s i den andre vil også kunne tilbys.) Denne kapasiteten kan abonnenten tilordne kanaler slik han ønsker.

Selv om abonnentene har mulighet til å benytte et utrolig antall kanaler og har en svært høy bitrate tilgjengelig, er det ikke forutsetningen at alle abonnenter skal gjøre full bruk av dette. Hensikten har vært å sette et så høyt "tak" at flest mulig behov skulle kunne dekkes innenfor de rammer som settes av den teknologi vi har i dag. Tanken er at en slik skal kunne dra nytte av masseproduksjon av utstyr basert på veletablert komponentteknologi for å lage framtidsrettet nett-tilkobling som etter hvert også skal bli rimelig.

Riktignok vil B-ISDN kreve abonnentlinjer basert på optisk fiber – noe som vil bli kostbart å etablere. Men selv for å kunne tilby N-ISDN vil en omfattende utbedring av abonnentlinjer være nødvendig i mange land – inkludert Norge. Og også i dette tilfellet ser det ut til å kunne bli regningssvarende å benytte fiberteknologi helt eller delvis. I de sentrale deler av nettet økes allerede i dag kapasiteten til transmisjonsutstyret ved at fiberteknologi tas i bruk. B-ISDN gir på denne måte muligheter til å utnytte den ressurs som investeringen i fiber representerer, slik at abonnentene får en best mulig overføringstjeneste ut fra den komponentteknologi en har tilgjengelig.

4 Overføringskrav fra ulike media kan "garanteres"

For brukeren er det oftest viktig å ha garanti for at han vil få tilstrekkelig tjenestekvalitet når han starter mediasesjoner. Dersom f eks overføringen av en videosekvens skjemmes av stadige forstyrrelser fra overføringsnettet, vil det kunne være mer fordelaktig å utsette overføringen eller starte en mindre ressurskrevende sesjon.

For at brukeren skal kunne garanteres den overføringskvalitet han har behov for, er det en fordel at han inngår "kontrakt" med nettet om dette; og at bruker og nett overholder kontrakten. Mekanismene som sikrer at kontrakten blir overholdt i B-ISDN, er knyttet til oppretting av forbindelser.

Når en forbindelse skal opprettes, beskriver brukeren sine minimumskrav til informasjonsoverføringen til nettet (f eks overføringskapasitet og maksimal variasjon i forsinkelsen, jitter). Nettet undersøker om det har tilstrekkelige ressurser til dette – sett på bakgrunn av eksisterende forbindelseskontrakter – og beslaglegger de ressurser som trengs i nettet for at den tjenestekvalitet som brukeren har bedt om, kan sikres mens forbindelsen opprettholdes. Dersom netttet ikke har de nødvendige ressurser tilgjengelig (f eks på grunn av at de er tilordnet andre forbindelser) avvises forespørselen. Brukeren får i så fall beskjed om dette og kan ta stilling til om en mindre krevende overføringskvalitet kan benyttes. (Nettet bør naturlig nok dimensjoneres så rikelig at slike avslag ikke må gis for ofte – innen akseptable kostnadsrammer.)

Det er ikke meningen at den enkelte *person* skal ha detaljert kunnskap om sitt overføringsbehov. Ved bestilling av forbindelser som er direkte styrt av personer, kan det være aktuelt å benytte standardiserte bærertjenester for media (f eks Video A – hvor A angir en av et begrenset antall video-overføringskvaliteter). Men i de fleste tilfeller vil nok informasjon om overføringsbehov lagres som en del av programvaren i terminalen for den enkelte applikasjon – kanskje da beskrevet ved minimumskravene for generelle overføringsparametre.

5 Taksering kan tilpasses krav fra ulike media

Ut fra den informasjon nettet får fra brukeren om de krav han stiller til de enkelte informasjonskanaler, reserverer nettet nødvendige ressurser til forbindelsene som opprettes gjennom nettet (f eks overføringskapasitet på den enkelte link og bufferplass i svitsjenodene). Det er denne ressursreservasjonen kunden bør betale for. For selv om kunden lar være å benytte ressursene i den grad han har beskrevet, vil nettet måtte holde dem i beredskap for han.

B-ISDN gir likevel anledning til å redusere den totale overføringskapasitet som må beslaglegges i nettet, ved å dra nytte av at informasjonsstrømmen på den enkelte kanal ikke er den samme hele tiden. (Ved kodet video f eks kan informasjonsstrømmen reduseres når bildet ikke endres særlig.) Forutsetningen er at kunden indikerer hvordan informasjonsstrømmen vil variere.

Kanaler fra ulike brukere som er i "aktiv" periode kan internt i nettet benytte overføringskapasitet fra kanaler som er i "rolig" periode. Informasjonsoverføringen kan slik gjøres rimeligere for kundene. Nettet kan spare ressurser når brukerens overføringsbehov varierer ved enten å benytte statistisk multipleksing eller ved aktivering/deaktivering av kanalen.

Statistisk multipleksing kan benyttes i deler av nettet som formidler så mange kanaler at en ut fra statistiske beregninger kan fastslå at sannsynligheten for at alle kanaler skal topp-belaste utstyret samtidig er så liten at dette kan ses bort fra. (Kanalene bruker her samme ressurs – f eks det fysiske medium i en overføringslink.) Den enkelte kanal kan da tildeles mindre kapasitet enn topp-belastningen tilsier – men kanskje noe mer enn en oppgitt gjennomsnittsbelastning for kanalen.

Første versjon av B-ISDNs bruker/nett-signalering vil ikke gi kunden anledning til å beskrive informasjonsoverføringsbehovet fullstendig. Brukeren kan med den bare oppgi topp-belastningen – og nettet må reservere ressurser i henhold til dette. Det vil da være prisgunstig for kunden å utjevne informasjonsstrømmen i tid mest mulig. Han vil da redusere toppbelastningen – og kan klare seg med en mindre ressurskrevende – og billigere – kanal.

Ved *aktivering/deaktivering* gir brukeren beskjed til nettet når kanalen ikke skal formidle trafikk for en periode. Nettet kan da disponere overføringskapasiteten til andre forbindelser inntil kanalen aktiveres igjen. Dersom nettet er avhengig av "tilfeldigvis" å ha tilgjengelig kapasitet når kanalen skal aktiveres, vil kunden kunne risikere en uakseptabelt lang venting før nettet kan etterkomme aktiveringsforespørsel. Benyttes derimot ressursene til overføring av trafikk med lavere prioritet (da helst trafikk med lave sanntidskrav), vil nettet kunne overføre kapasiteten tilbake til den deaktiverte kanalen straks brukeren ber nettet om å få den aktivert igjen.

I motsetning til aktivering/deaktivering kan nettoperatøren bruke statistisk multipleksing uten at brukerens utstyr er involvert i dette. Nettoperatøren kan slik endre sin ressurstildelingsstrategi uten at brukeren behøver å gjøre noen tilpasning på sin side – bare han på forhånd oppgir hvilken toppbelastning og hvilken gjennomsnittlig belastning kanalen må kunne tåle.

Fordelen med aktivering og deaktivering framfor statistisk multipleksing er at metoden kan benyttes selv når trafikkendringene på kanalen er store sett i forhold til totaltrafikken, og en er heller ikke avhengig av at mange kanaler benytter fellesressursen. Teknikken vil bl a være nyttig i de tilfeller hvor brukeren på forhånd vet at han under multimedia-sesjonen vil få bruk for informasjonskanaler på et senere tidspunkt – og at det har liten hensikt å starte multimediasesjonen uten dem. Han kan da reservere kanaler slik at han er sikker på å ha dem tilgjengelig og holde dem deaktivert inntil han får bruk for dem – for slik å redusere sine overføringsutgifter. Avhengig av situasjonen vil derfor både statistisk multipleksing og aktivering/deaktivering kunne benyttes – gjerne i kombinasjon – for å utnytte nettressursene og gi abonnentene rimeligere informasjonsoverføring.

Fordelen med aktivering/deaktivering av en og samme kanal i stedet for å opprette og nedkoble kanaler etter behov, er at brukeren kan få overføringskapasitet raskt tilgjengelig. Men for at kunden skal være *sikker* på å få aktivisert kanalen raskt, må som nevnt kanalen ha prioritet. Trafikkmessig vil det være lite å vinne på å la kanalen ha tilordnet sin kapasitet i deaktivert tilstand.

Det er foreløpig noe uklart om aktivering/deaktivering kan skje raskere med signalering til nettet gjennom kanalene som overfører informasjon mellom brukerne (refereres til ved den engelske termen "fast reservation") eller om en like gjerne kan benytte de aktiverings-/deaktiveringsmekanismer man uansett vil ha i signaleringskanalen. Kanskje vil det da også bli enklere å tildele deaktiverte ressurser til lavprioritets kanaler siden en signalerer direkte mot mer sentrale deler av signaleringsnettet.

6 Multimedia informasjon kan overføres på en eller flere kanaler

Selv om B-ISDN gir muligheter for å tilpasse kanalene til de behov som de enkelte media stiller, vil det likevel være et kostnadsspørsmål om det er mest hensiktsmessig for kunden å opprette flere kanaler gjennom det offentlige nettet mellom to multimedia-stasjoner, eller om alle enkeltmedia skal kommunisere over en og samme overføringskanal. Mange ressurser som nettet må tilordne kanalene, er helt eller delvis uavhengig av overføringsrate og overføringskvalitet (f eks plass i rutingtabeller eller utstyr for å overvåke at kontrakten som er inngått om kanalen overholdes, politifunksjoner).

Riktignok er det definert to multipleksnivåer for B-ISDN. Abonnenten kan velge mellom 256 kanalgrupper (virtual paths), hver med 65 536 kanaler. Men det har ikke vært meningen at abonnentene skulle beslaglegge en hel kanalgruppe for en multimediaforbindelse. Selv om kanalene i en gruppe kan beslaglegges med kun en oppførelse i den enkelte tabell og politifunksjonsenheter som dekker hele kanalgruppen, vil fort en så omfattende beslaglegging av kanaler bli for kostbar. Dette fordi det beslaglegges et svært høyt antall kanaler langs hele forbindelsen, og antall kanaler tross alt vil være begrenset på den enkelte fysiske link i nettet.

Dersom flere media skal benytte samme kanal må en derfor kunne multiplekse/demultiplekse de ulike media-strømmene. Foreløpig er det ikke helt klart hvordan en slik multipleksing/demultipleksing skal skje i B-ISDN. I arbeidet som nå pågår med å utforme B-ISDN er det sterke krefter som ønsker eventuelt å skyve slik multipleksing opp til høyere protokoll-lag (over de karakteristiske B-ISDN informasjonsoverføringsprotokollene (over adapsjonslaget)) – dersom multipleksingen ikke kan skje i ATM-laget.

7 Forbindelsesfri kommunikasjon i B-ISDN kan dekke noen av kravene som multimedia stiller

Som allerede nevnt er overføringsmekanismen i B-ISDN (ATM) forbindelsesorientert. Før informasjon kan overføres må det etableres en overføringskanal gjennom nettet. Dette tar tid og involverer signaleringsressurser i terminaler og nett. Opprettede forbindelser vil også legge beslag på ressurser i nettet. Selv deaktiverte forbindelser krever at noen ressurser er beslaglagt.

Nøyaktig hvor lang tid det vil ta å opprette en forbindelse i B-ISDN og hvor ressurskrevende/kostbart dette vil bli, er usikkert. I ISDN regner en med en oppkoblingstid på i størrelsesorden 1 sek. En må forvente at B-ISDN vil tilby avanserte overføringstjenester som vil kreve mye prosessering. Spørsmålet er om enkle og hurtige forbindelsestjenester – slik som i Datex med en oppkoblingstid i størrelsesorden 10 msek – også vil bli tilbudt. På sikt vil den teknologiske utviklingen gjøre at selv avanserte overføringstjenester vil kunne effektueres raskt.

Selv om den grunnleggende overføringsmekanisme i B-ISDN er forbindelsesorientert, kan B-ISDN slik som N-ISDN kompletteres med forbindelsesfri servere. En egnet forbindelsesfri nettprotokoll for B-ISDN er allerede under utvikling. Abonnenter kan opprette permanente eller semipermanente forbindelser mot disse og via dem komme i kontakt med andre brukere som også kommuniserer forbindelsesfritt. (Direkt samtrafikk mellom brukere som benytter forbindelsesfri og forbindelsesorientert kommunikasjonstjenester har vist seg å bli lite brukt i eksisterende nett.)

Forbindelsesfri kommunikasjon vil teknisk sett være et bedre alternativ enn forbindelsesorientert kommunikasjon dersom alle tre betingelsene nedenfor er oppfylt *samtidig:*

- Terminalene kommuniserer relativt sjeldent med hverandre (ellers kan det lønne seg å benytte en periodevis aktivert forbindelse)
- Bare lite informasjon overføres av gangen (ellers kan en sette opp forbindelsen for den tiden en trenger på overføringen)
- Det ikke er kritisk om informasjon blir borte underveis, stokkes, ikke overføres korrekt – eller har svake tidskrav (ellers kan det være lønnsomt å benytte den tjenestegaranti en kan oppnå ved forbindelsesorientering).

Men ofte vil andre forhold enn de rent tekniske avgjøre hvorvidt en skal overføre forbindelsesfritt eller ikke. Ikke minst viktig er hvilke protokoller de kommuniserende parter har tilgjengelig. Prisen en må betale for anskaffelse av kommunikasjonsutstyr – og ikke minst kostnadene for overføringen – betyr mye dersom en er i den situasjon at en kan velge overføringsmetode.

Særlig i en innledende fase kan forbindelsesfri kommunikasjon bli et interessant alternativ i B-ISDN. Dette fordi det signaleringssystem som i B-ISDN skal danne grunnlag for forbindelsesorientert kommunikasjon, ikke er ferdig utviklet ennå (første versjon av standarden forventes ferdig i 1992). Den forbindelsesfrie nettprotokollen er som nevnt heller ikke ferdig utviklet, men blir sannsynligvis enklere – og vil derfor kanskje komme raskere som produkt. Men sterke tidskrav som enkelte media setter (f eks bildeoverføring ved konferanser), medfører store krav til at nettet har tilstrekkelige ressurser tilgjengelig mens kommunikasjonen pågår. Det vil antakelig bli problematisk å basere slik kommunikasjon bare på forbindelsesfri informasjonsoverføring.

8 Synkrone multimedia-applikasjoner kan overføres asynkront

I B-ISDN er ikke bare den forbindelsesfri – men også den forbindelsesorienterte trafikk asynkron. Sammen med hvert datafragment som sendes gjennom nettet sendes det også informasjon om hvor fragmentet skal sendes videre. Denne teknikken er svært forskjellig fra den teknikk som benyttes i telefonnettet og i N-ISDN. Her kommer informasjon som skal overføres i en bestemt retning, inn til svitsjepunkter i nettet i faste periodevise tidsintervall. Hele telefonnettet krever derfor en sterk synkronisering for å sikre at riktig informasjon ankommer riktig sted til riktig tid.

Siden en ikke har behov for denne sterke synkroniseringen for å få til informasjonsoverføringen i B-ISDN, kan det være aktuelt å gjøre nettet mer asynkront. (Synkronisering på den enkelte link kan likevel være aktuelt – slik at sender og mottaker kan benytte en felles klokke for å avgjøre når et bit som overføres begynner og slutter. Det kan også være aktuelt å benytte transmisjonssystemer med enda høyere grad av synkronisering.)

Sanntids overføring av video og tale krever at sender og mottaker er synkroniserte. En mottaker må kontinuerlig presentere lyd/bilde for brukeren med samme takt som senderen benytter.

Dersom den presenterer informasjonen for sakte, blir den hengende stadig lengre etter – inntil informasjonsbufrene flyter over. Dersom den presenterer informasjonen for fort, vil den i visse tidsintervaller ikke ha informasjon å presentere.

Dersom nett-tilknytningene ikke har en felles synkronisering vil ikke nettet heller kunne benyttes til å synkronisere terminalene. Terminalene må da selv ha mekanismer for synkronisering. (En aktuell algoritme er at en mottaker overvåker informasjonskøen og justerer presentasjonstakten etter hvor mye informasjon som står i bufferet.)

9 Brukeren kan få tilgang til avanserte nett-tjenester i B-ISDN

Det å etablere forbindelsen gjennom nettet er bare en av komponentene som inngår i en overføringstjeneste. Andre viktige komponenter er å taksere, undersøke om abonnenten kan ta imot anropet, varsle abonnenten om at nytt anrop kommer. I de overføringstjenester (bærertjenester) som tilbys i dag, er alle disse komponenter i tjenesten knyttet til det å sette opp en (toveis) forbindelse.

En multimediasesjon vil bestå av flere enkelt-media sesjoner – kanskje hver av dem med en eller flere forbindelser. En slik sammensatt sesjon krever en frikobling av andre komponenter fra etableringen av forbindelsen. Det vil f eks ikke være naturlig å varsle brukeren om multimedia-sesjonen hver gang det settes opp en forbindelse til han. I B-ISDNs overføringstjenester ønsker en derfor å skjelne sterkere mellom anropsbehandlingen (som har relevans for hele multimedia-sesjonen) og opprettelsen av den enkelte forbindelse.

B-ISDN vil omfatte en rekke ytelser fra nettet som støtter opp om informasjonsoverføringen. I dagens nett tilbys slike ytelser som *tilleggstjenester*. Eksempler på dette er overføring av brukerens telefonnummer til en annen abonnentlinje (medflytting) og overføring av debitering til den som blir oppringt. Andre ytelser som er aktuelle i B-ISDN er koordinering av konferanser og mulighet for brukere til å overføre sitt abonnement til en hvilken som helst terminal (personlig kommunikasjon).

Stadig ny tilvekst til tjenesterepertoaret har kommet til i de senere år, og en er nå svært opptatt av å organisere telenett slik at nye ytelser kan introduseres raskt – og med investeringer som tilsvarer deres etterspørsel. En egen arkitektur for nettkontroll anbefales av CCITT for å oppnå dette. Den går under betegnelsen Intelligente Nett, IN.

Hvordan IN vil påvirke bruker/nett-signaleringen i B-ISDN er for tiden noe uklart. Ikke bare fordi alle ytelsene som IN skal dekke naturlig nok ikke er spesifisert ennå, men også fordi det er uklart hvordan informasjonsutvekslingen mellom bruker og nett skal organiseres i IN. IN ser i dag ut til å sette overføring og andre nettfunksjoner inn i samme rammeverk, men samtidig er det en målsetning at IN skal kunne benyttes også for eksisterende nett hvor skillet mellom overføringstjenester og andre ytelser er mer markant. Et viktig spørsmål er derfor om tradisjonell aksess-signalering for kontroll av forbindelser – og informasjonsoverføringen mellom bruker og nett for kontroll av det som i dag betraktes som tilleggstjenester, skal ses på som et integrert hele i B-ISDN.

Den utvidede fleksibilitet som brukerne får til å styre nettfunksjoner i B-ISDN vil nødvendigvis måtte føre til et ganske *omfattende* signaleringssystem. Det vil være viktig å konstruere det slik at det i tillegg ikke blir unødig *innfløkt*. En enkel og ryddig meldingsstruktur synes derfor å være viktig å oppnå – gjerne tilpasset den enkelte brukerens behov. Når brukeren er et menneske, vil det bli viktig at B-ISDN terminalene støtter opp om brukerens nettkontroll slik at styring av kompliserte nettfunksjoner forenkles (f eks med menyvalg, ikoner, bilder og talte instruksjoner).

10 Konklusjon

B-ISDN er et nettkonsept for multimedia som kan benyttes i både offentlige og private nett. Konseptet setter høyt tak for de krav multimedia-applikasjoner kan stille, men gir likevel gode muligheter for å tilpasse overføringsressursene til det behov den enkelte bruker har. Siden B-ISDN gir en enhetlig løsning på et allsidig overføringsbehov, vil konseptet på sikt trolig kunne gi en rimelig nettløsning for multimedia-overføring.

Brukerne inngår avtale med nettleverandøren om den enkelte overføringskanal (forbindelsesorientering). Det kan derfor gis garanti for at informasjonen kommer riktig fram under mediasesjonene. Dette er særlig viktig for multimediaoverføring med sterke tidskrav – f eks videokonferanser.

Mekanismer som statistisk multipleksing og aktivering/-deaktivering av kanaler gjør at overføringsressurser i nettet kan utnyttes optimalt, selv om brukeren er garantert tilgang til nettressursene når han trenger dem. (Ved deaktivering må ressursene i så fall tilordnes kanaler med lavere prioritet.)

Forbindelsesfri kommunikasjon vil også kunne tilbys i B-ISDN. Slik kommunikasjon vil være godt egnet for informasjonsoverføring med svake tidskrav (f eks transport av informasjon mellom lagringsenheter i elektroniske multimedia meldingssystemer). Særlig i en innledende fase kan forbindelsesfri informasjonsoverføring bli svært viktig.

Et B-ISDN basert nett behøver ikke å være helt synkront, men ved visse applikasjoner vil da terminalene måtte ha mekanismer som sørger for at sendere og mottakere sikres innbyrdes synkronisering.

Brukeren vil få tilgang til et omfattende spekter av ytelser i B-ISDN, og organiseringen av kontrollfunksjonene vil kunne gi brukere rask adgang til et tjenesterepertoar i sterk utvikling – forhåpentligvis uten at signaleringssystemet blir unødig komplisert. Det vil likevel være en viktig oppgave for utviklere av de framtidige B-ISDN terminaler at brukerne kan få tilgang til ytelsene på en enkel og oversiktlig måte.

Broadband Integrated Services Digital Network

BY ERIK SKÅTUN

621.39.05

Introduction

Broadband Integrated Services Digital Network, B-ISDN, will be developed into a network concept for transmission of SOUND, VIDEO AND DATA incorporating bit-rates from 100 bits/second to over 100 Mb/second.

B-ISDN services will be divided into two main groups:

- Interactive services requiring two-way communication, and
- Distribution services including traditional broadcasting.

Evolving technologies together with demand for high-speed communications constitute the driving forces behind the new network evolution. Obviously new high capacity network highways are required for transport of information at these speeds,

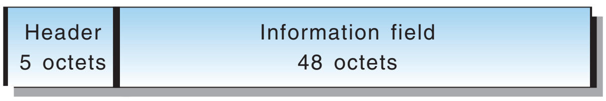


Figure 1 ATM cell structure

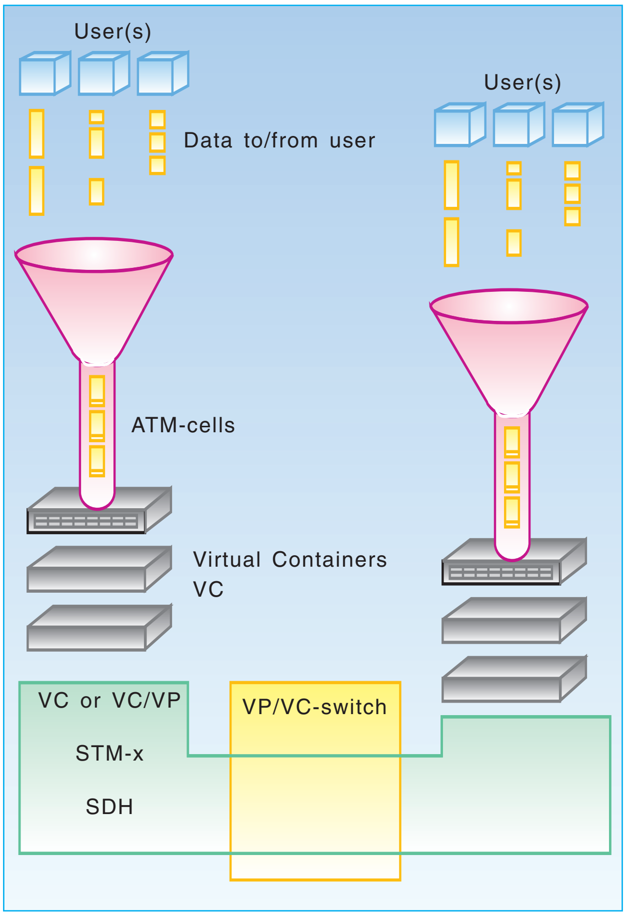


Figure 2 Note: VC denotes Virtual Container, the package used for "wrapping" ATM cells into the transmission medium. In the diagram VC also denotes Virtual Channel, not to be confused with Virtual Container

the SDH, Synchronous Digital Hierarchy, will provide such "motorways". The SDH has been standardised by CCITT, and will provide flexible multiplexing of varied rate of information flow together with advanced capabilities for operations, administrations, and maintenance of the broadband network. The CCITT has (in Rec. G.707) recommended 3 levels in the hierarchy: STM-1 (155,520 Mbit/s), STM-4 (622,080 Mbit/s) and STM-16 (2.488320 Gbit/s), higher rates to be recommended when technology allows.

Recommendation CCITT I.121 states that "Asynchronous transfer mode, ATM, is the transfer mode for implementing B-ISDN and is independent of the means of transport at the Physical Layer". ATM is based upon a method of transporting information in CELLS of fixed length. The cells comprise 48 octets of "payload" with a "header" of 5 octets added, the header containing the cell identification and routing information. Cells are injected at the A-side edge of the network, transported in "containers" in the SDH and extracted at the destination edge for delivery to the B-side.

The ATM-header identifies the ATM-cell as belonging to a "virtual path" and "virtual channel" and thus describes the "road" the ATM-cell is to be passed along on its way to the destination. Cells with identical header (on the same section) belong to the same "connection".

ATM-cells will be transported in "containers" which again are mapped into the STM-1 structure of the SDH. For European standards, following containers are specified: C12 (2,048 Mbit/s), C3 (34,368 Mbit/s), and C4 (139,264 Mbit/s). In one STM-1 frame there is room for 1 VC-4, or 3 VC-3, or 63 VC-12. (VC = Virtual Container). A VC-4 container is shown in figure 3.

The VC-4 container has "path overhead" bytes, the bytes shown left in the diagram, 9 bytes in all. The byte J1 is the first byte of the container. 260 x 9 bytes are available for the ATM cells and required "stuffing" cells. This container is then mapped into an STM-1 as payload. It follows that one VC-4 container could make up the entire payload of one STM-1, however, there is no synchronism between containers and STM implying that a VC-4 may be positioned anywhere in the rows/columns of an STM-1. The beginning of a VC-4 is, however, marked with the J-byte.

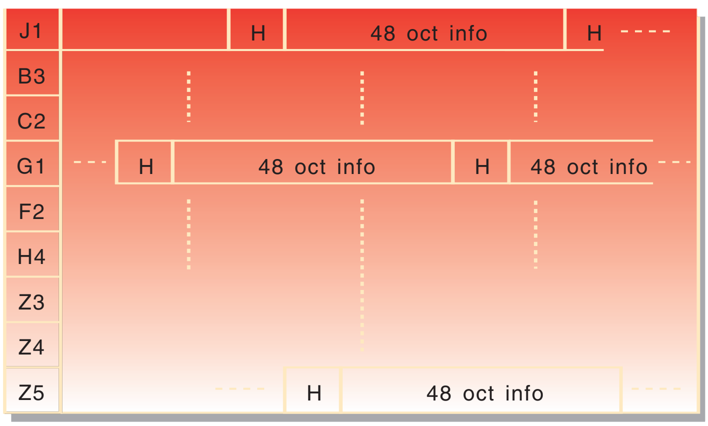


Figure 3 A VC-4 container

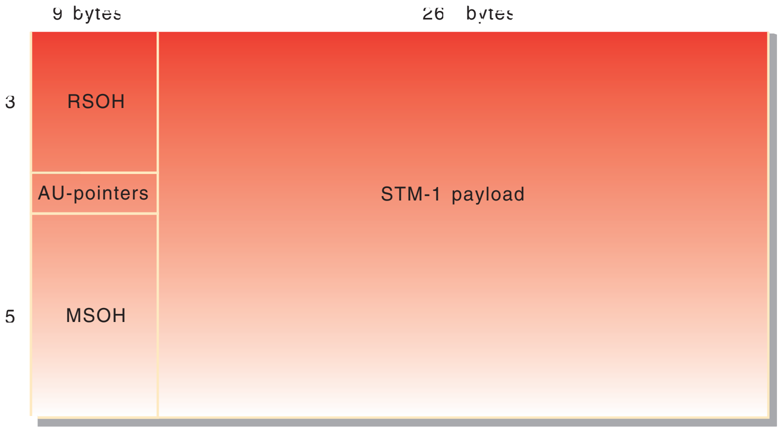


Figure 4 Diagram for a STM-1 frame

A1	A1	A1	A2	A2	A2	C1	nu	nu	
B1	ms	ms	E1	un	un	F1	nu	nu	
D1	ms	ms	D2	un	un	D3	nu	nu	
AU-pointers									
B2	B2	B2	K1	un	un	K2	un	un	STM-1-
D4	un	un	D5	un	un	D6	un	un	payload
D7	un	un	D8	un	un	D9	un	un	
10	un	un	11	un	un	12	un	un	
Z1	Z1	Z1	Z2	Z2	Z2	E2	nu	nu	

Figure 5 STM-1 section overhead (10, 11 and 12 = D10, D11, D12)

The position of this J-byte will be identified by the AU-pointers in figure 4 (the bytes preceding the J-byte belonging to another C-4 container with its remaining information in the previous STM-1 frame). The diagram for an STM-1 frame is shown in figure 4.

An STM-1 is either a sole individual element of a transmission section or it is multiplexed into higher order STM. The section overhead bytes is divided into two parts: The ROSH = regenerator section overhead and the MSOH = the multiplexer section overhead, in addition to the AU-pointers. The significance of the various bytes are:

A1, A2: Frame sync.
B1, B2: Fault supervision
C1: STM identifier
D1-D12: Data channels
E1, E2: Service channels, speech
F1: User channel
K1, K2: Automatic protection switching. Some channels are reserved for national use, nu, others are unused = un or spare Z1, Z2.

B-ISDN Architecture

The major characteristics of ATM are:

- Use of fixed length cell for information transfer, 48 octets of length

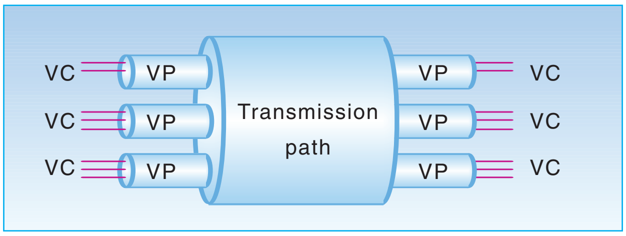


Figure 6 Relations between transmission path, virtual path and virtual channel

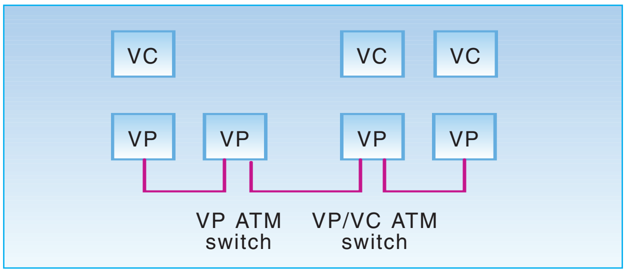


Figure 7 Virtual path switching

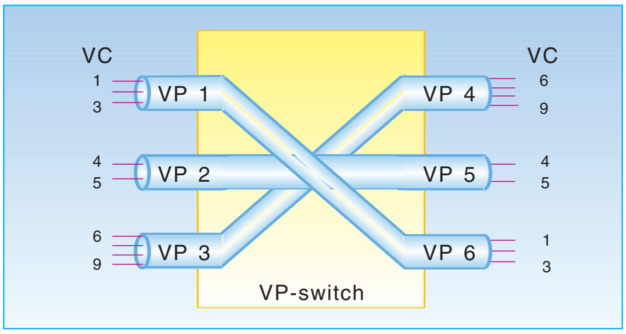


Figure 8 Virtual channel/virtual path switching

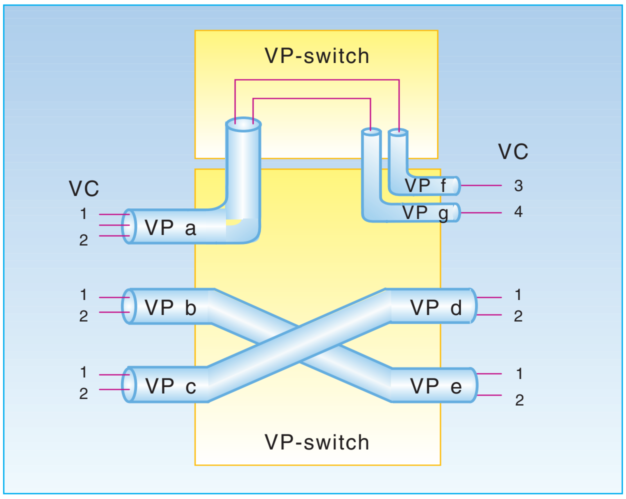


Figure 9 Virtual channel/virtual path switching

- Identification of each communication unit by cell header, 5 octets long.

An ATM label in the header is structured into two fields for identification of virtual channel (VC) and virtual path (VP).

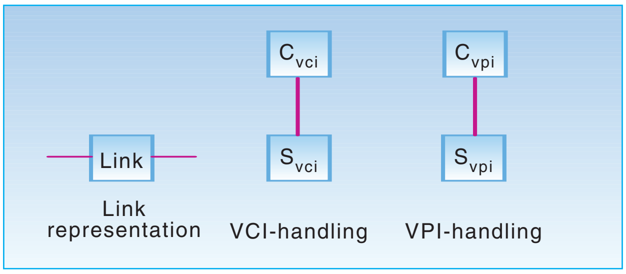


Figure 10 General connection element model

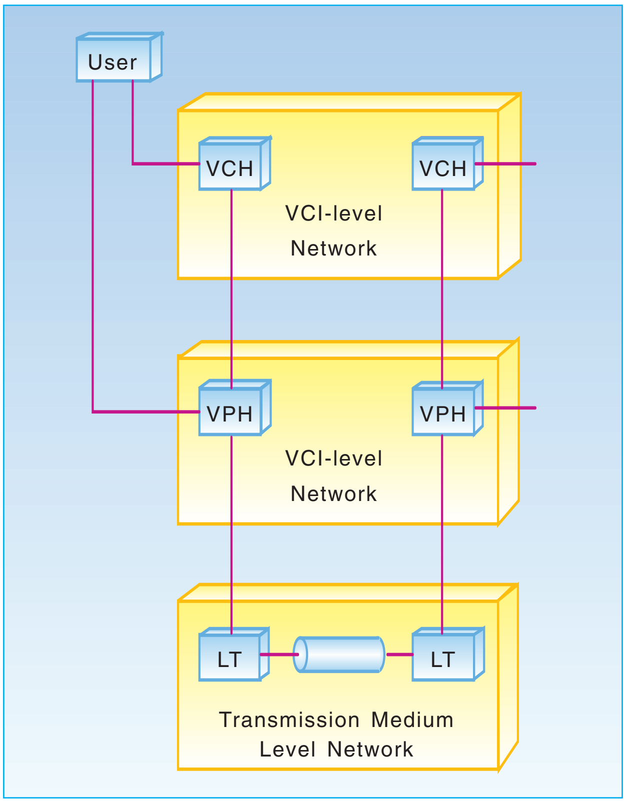


Figure 11 ATM-based B-ISDN architecture

In B-ISDN, the virtual path connection is introduced for routing groups of virtual channels in the network. Therefore two levels of connection handling will exist in the B-ISDN. These levels must be represented by two different switching blocks in the connection elements, one switching according to the virtual path identifier (VPI) the other switching according to the virtual channel identifier (VCI), each switching block under control of its respective control function.

A general connection element model in B-ISDN is described using five functional blocks: Switching block for VPI, control block for VPI; switching block for VCI, control block for VCI and an interconnection link. The link block incorporates all the functions implementing the physical layer. Different links may be identified, e.g. access links and transit links.

Figure 11 shows the ATM-based B-ISDN architecture. It is seen that the transport network is structured in three levels: VC level, VP level and transmission medium level. VC level connections are typically established and released on customer demand basis.

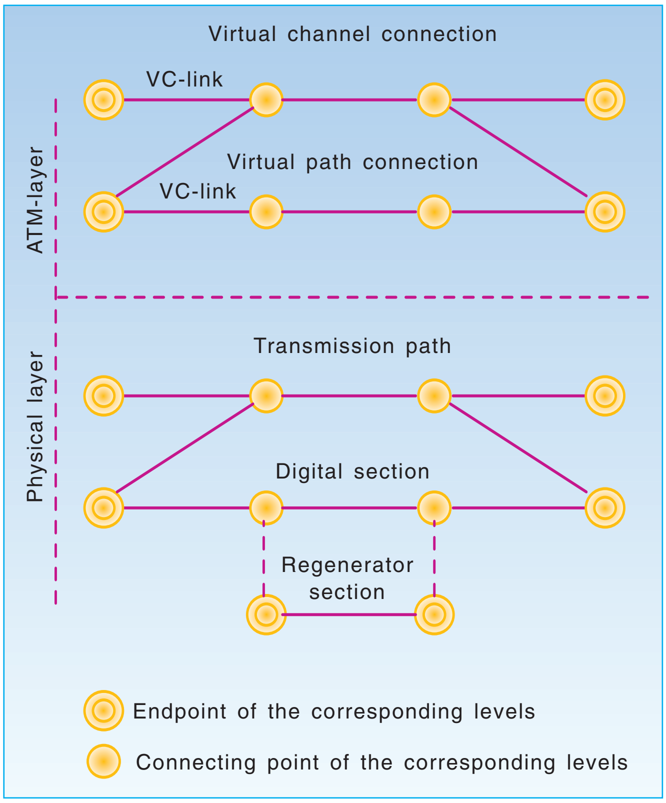


Figure 12 Layers in the hierarchy

Hierarchical layer-to-layer relationship

Links are concatenated to form a connection, defined between two end-points. For example, terminal equipment corresponds to the end-point in the VC level connection, whilst a VC switch corresponds to the end-point in the VP level connection in addition to forming a connecting point in the VC level connection.

A virtual channel identifier (VCI) identifies a particular VC link in a given VP connection. A VC switch translates the VCI values of incoming VC links into VCI values of the outgoing VC links. VC links are concatenated to form VC connection (VCC). At the VCC end-point, the cell information field is exchanged between the ATM layer and the user of the ATM layer service. At the VC connecting point, the VPC supporting the incoming VC links are terminated and a new VPC is created. The ATM layer provides cell sequence integrity for cells belonging to the same VCC. In the same manner, a virtual path identifier (VPI) identifies a collection of VC links sharing the same VPC. A VP link is originated or terminated by assigning or removing VPI values. In a VP connecting point the VPI values of incoming VP links are translated into VPI values of the outgoing VP link. Also VP links are concatenated to form a VPC. The physical layer is subdivided into three levels: the transmission path level, digital section layer and regenerator section level.

Protocol reference model

The B-ISDN Protocol Reference Model is composed of a user plane, a control plane and a management plane. Above the Physical Layer, the ATM Layer provides call transfer for all services and the AAL implements service-dependent functions

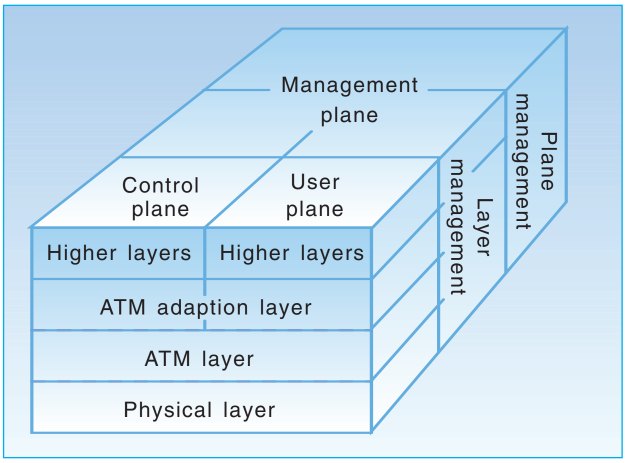


Figure 13 B-ISDN Protocol Reference Model

Layer management	Functions		
	Higher layer functions	Higher layers	
	Convergence	CS	AAL
	Segmentation and reassembly	SAR	
	Generic flow control Cell header generation/extraction Cell VPI/VCI translation Cell multiplex/demultiplex	ATM	
	Cell rate decoupling HEC header sequence generation/verific Cell delineation Transmission frame adaption Transmission frame generation/recovery	TC	Physical layer
	BIT timing Physical medium	PM	

Figure 14 Layers of the PRM (Protocol Reference Model) and functions of the Physical layer, the ATM layer and the ATM Adaption layer (AAL). CS = convergence sublayer, PM = physical medium, SAR = segmentation and reassembly sublayer, TC = transmission convergence

8	7	6	5	4	3	2	1	bit / oct
GFC				VPI				1
VPI				VCI				2
VCI								3
VCI				PT		RES	CLP	4
HEC								5

Figure 15 Cell structure at UNI, User Network Interface

8	7	6	5	4	3	2	1	bit / oct
VPI								1
VPI				VCI				2
VCI								3
VCI				PT		RES	CLP	4
HEC								5

Figure 16 Cell structure at the NNI, Network Node Interface

to the layer above the AAL. The layer above the AAL in the control plane performs call control and connection control. The management plane provides network supervision functions.

User plane: The user plane provides for transfer of user information, along with associated controls such as flow control and recovery from errors, etc.

Control plane: The control plane performs the call control and connection control functions; undertaking the signalling required to establish, supervise and release calls and connections.

Management plane: The management plane provides two types of functions, namely the plane management functions and the layer management functions. The plane management performs management functions related to a system as a whole and provides co-ordination between all the planes. Plane management has no layered structure.

Layer management performs management functions, including META-SIGNALLING, relating to resources and parameters residing in its protocol entities. Layer management also handles the OAM-information flows specific to the layer concerned.

Functions of individual layers

Some terms:

Idle cell (physical layer): Cell inserted/extracted by the physical layer in order to adapt the cell flow rate at the boundary between the ATM-layer and the Physical-layer to the available payload capacity of the transmission system used.

Valid cell: Cell whose header has no errors, or has been modified by the HEC, the header error control verification process (errors have been corrected).

Invalid cell: Cell whose header contains errors not dealt with by the HEC. Such cells are discarded at the physical layer.

Assigned cell (ATM-Layer): Cell providing service to an application using the ATM-layer service.

Unassigned cell is the opposite of assigned cell.

The Physical layer consists of two sublayers. The physical medium, PM sublayer includes only physical medium dependent functions. The transmission convergence (TC) sublayer performs all functions required to transform a flow of cells into a flow of data units (bits) which can be transmitted and received over a physical medium. The service data unit (SDU) crossing the boundary between the ATM-layer and the Physical layer is a flow of valid cells. The data flow inserted into the transmission system payload is physical medium independent and self-supported.

Cell structure coding

GFC (UNI only), generic flow control is intended to provide the ATM layer with a multipoint control function. GFC does not appear on the NNI.

VPI/VCI is the routing field of the header. The VPI at UNI consists of 8 bits, at the NNI 12 bits. VCI comprises 2 octets at NNI and UNI. The VPI/VCI uniquely identifies the “call”.

PT identifies the Payload Type, default value 00 for user information.

RES is reserve, 1 bit.

CLP is a bit for Cell Loss Priority, value 0 being highest priority.

HEC is an 8-bit field for Header Error Control

HEC is constructed performing a polynom-division-operation on the first 4 octets of the header, presenting the calculated syndrome in the 5th octet. This is the method of header error control. This is also used to identify the beginning of an ATM cell. No error control of the information field is performed by the network.

B-ISDN ATM Adaption Layer, AAL

The ATM adaption layer enhances the services provided by the ATM layer to support the functions required by the next higher layer. The AAL performs functions required by the user, control and management planes and supports the mapping between the ATM layer and the next higher layer. The functions performed in the AAL depend upon the higher layer requirements. The AAL supports multiple protocols to fit the needs of different AAL service users. The AAL is therefore service dependent.

The AAL isolates the higher layers from the specific characteristics of the ATM layer by mapping the higher layer protocol data units (PDUs) into the information field of the ATM cell and vice versa.

The AAL has two sublayers: Segmentation and Reassembly sublayer, SAR, and the convergence sublayer, CS.

SAR: Performs the functions of segmentation of higher layer information into a size suitable for the ATM cell structure; and the reverse: reassembling the contents of ATM cells information fields into higher layer information.

CS: The prime function is to provide the AAL service at the AAL-SAP (Service access point). This sublayer is service dependent.

Examples of service classes:

Class A: Circuit emulation; constant rate video

	Class A	Class B	Class C	Class D
Timing relation between source and destination	Required		Not required	
Bit rate	Constant	Variable		
Connection mode	Connection oriented			Connection less

Figure 17 Service classification for AAL

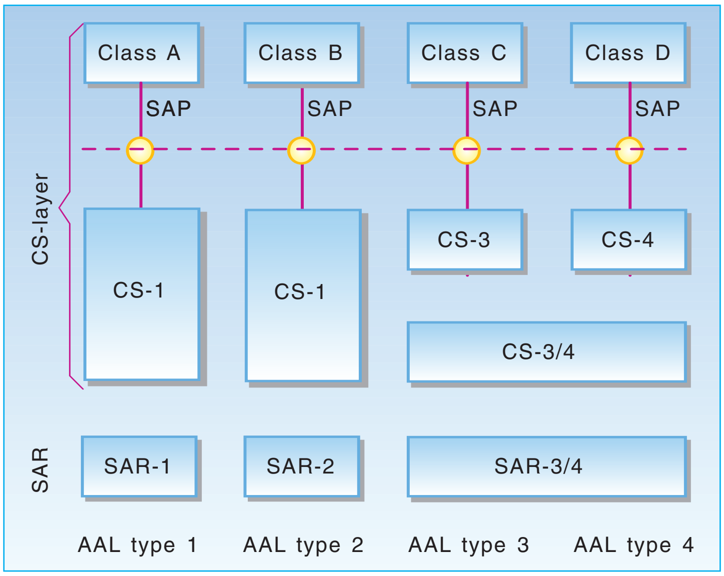


Figure 18 Relationship between the service classes and the layers of AAL

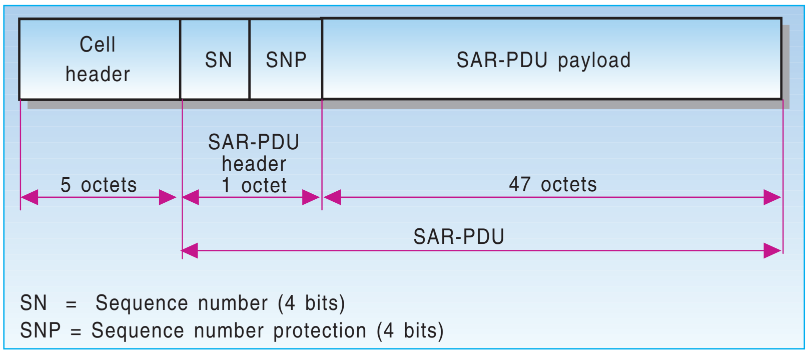


Figure 19 SAR-PDU format for AAL type 1

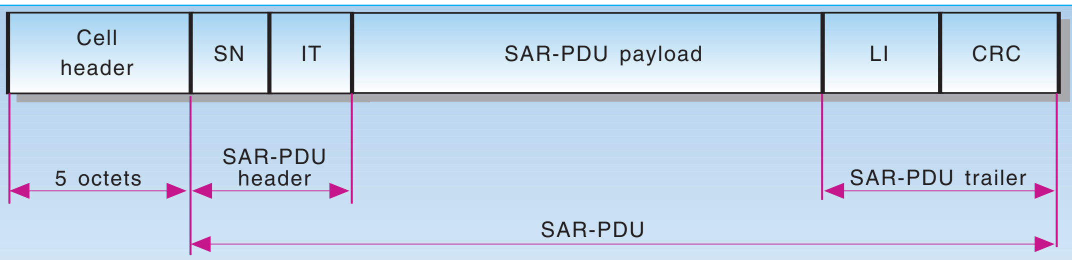
Class B: Variable bit-rate video and audio

Class C: Connection-oriented data transfer

Class D: Connectionless data transfer

The relationship between the service classes and the layers of AAL is shown in figure 18.

The AAL receives from the ATM layer information in the form of a 48 octet ATM Service Data Unit, ATM-SDU, and the reverse: AAL passes information in the form a 48 octet ATM-SDU to the ATM layer.



SN = Sequence number
IT = information type used to indicate beginning of message (BOM), continuation of message (COM), end of message (EOM), timing information and also component of the video or audio signal
LI = length indicator
CRC = cyclic redundancy check code

Figure 20 Example of SAR-PDU format for AAL type 2. SN = sequence number, IT = information type used to indicate beginning of message (BOM), continuation of message (COM), end of message (EOM), timing information and also component of the video or audio signal, LI = length indicator, CRC = cyclic redundancy check code

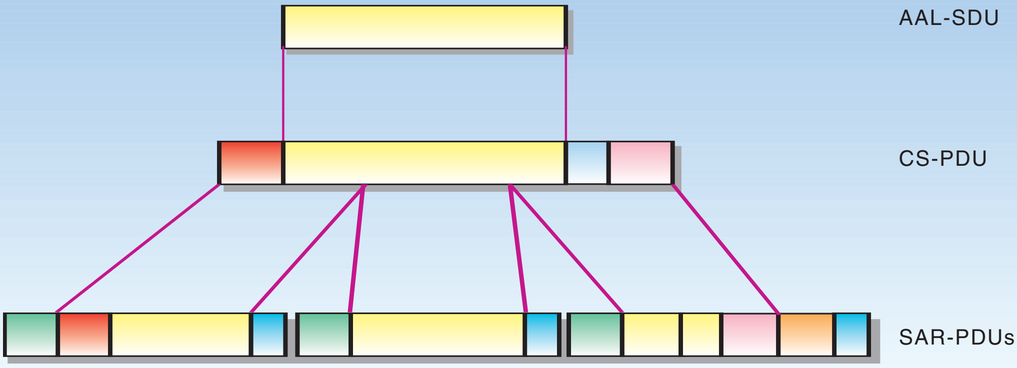


Figure 21 Message Mode Service

AAL type 1

The services provided by AAL type 1 to the higher layer are:

- Transfer of service data units with a constant source bitrate and delivery with the same bitrate
- Transfer of timing information between source and destination
- Indication of lost or erred information not recovered by AAL type 1.

The functions to be performed in the AAL include: segmentation and reassembly of user information, handling of cell delay variation, handling of lost and misinserted cells, recovery of source clock at the receiver and miscellaneous error monitoring functions. Segmentation and reassembly sublayer functions will be performed on an ATM-SDU basis.

AAL type 2

AAL type 2 has not been completely specified, and length and information of the control fields have not been determined.

AAL type 3

The AAL type 3 is used for connection oriented variable bitrate service. Two modes of service are defined: Message Mode Service and Streaming Mode Service.

The Message mode service supports transport of a single AAL-SDU in one (optionally more than one), CS-PDU.

The Streaming Mode Service provides transport of one or more fixed size AAL-SDU(s).The AAL-SDU may be as small as one octet, and is always delivered as a unit. Thus this service is suitable for the transfer of low speed data with low delay requirements.

SAR-PDU format for AAL type 3 (see figure 23):

ST	Segment type, 2 bit
SN	Sequence number, 4 bit
RES	Reserved field 10 bit
LI	Length indicator, 6 bit
CRC	Cyclic redundancy check code, 10 bit

Coding of segment type is:

Beginning of message, BOM:	10
Continuation of message, COM:	00
End of message, EOM:	01
Single segment message, SSM:	11

Thus 44 octets are left for SAR-PDU payload. This payload is fully utilised for BOM and COM, whilst EOM and SSM has variable length (Message mode), in the Streaming mode the payload of all segments depend on the AAL-SDUs. The length indicator, LI, determines the length in octets of the payload.

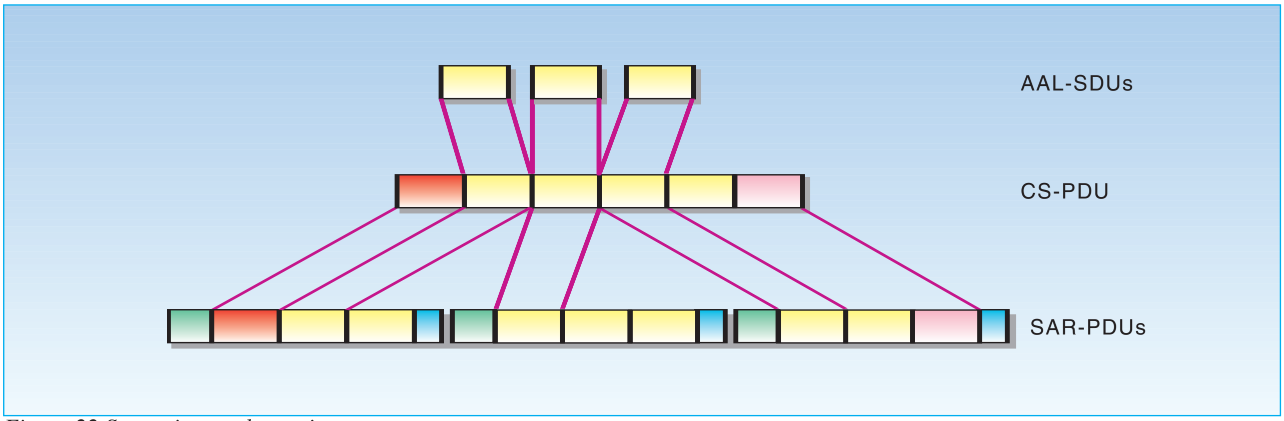


Figure 22 Streaming mode service

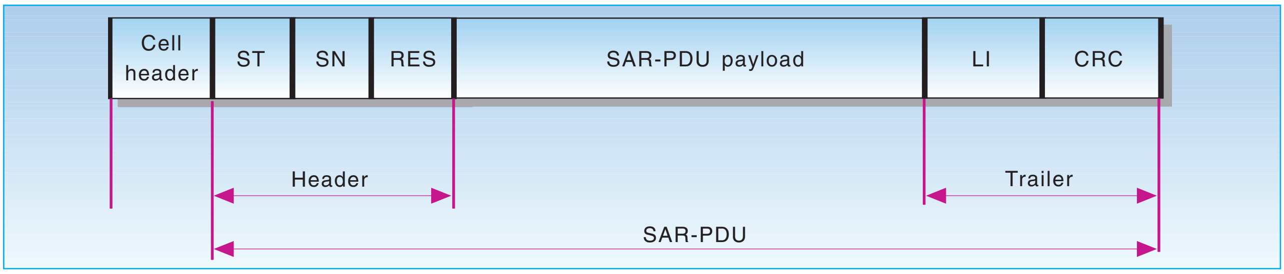


Figure 23 SAR-PDU format for AAL type 3

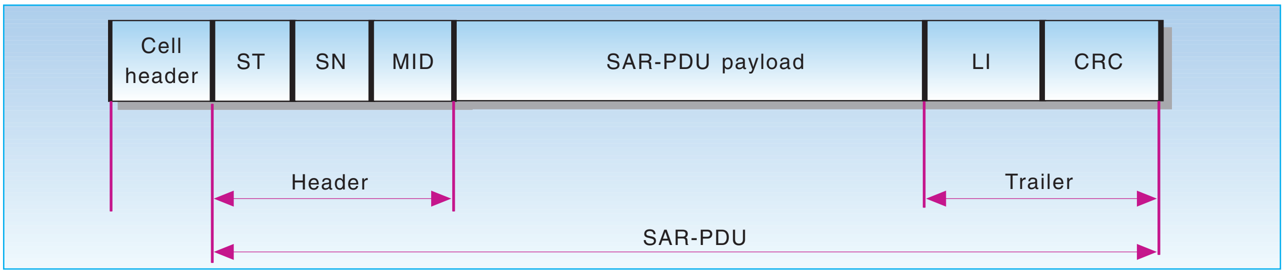


Figure 24 SAR-PDU format for AAL type 4

AAL type 4

The AAL type 4 provides the capabilities to transfer the AAL-SDU from one AAL-SAP to one or more AAL-SAPs through the ATM network with a QOS negotiated between user and network.

Protocol Overhead Information

The diagram in figure 25 shows that the ATM layer payload includes protocol information necessary for the layers above, thereby reducing the “net” payload by the required number of octets that serve higher layers requirements for such protocol overhead. As the protocol data units (PDU) are passed over the layer boundaries, the next layer will add its protocol information.

From N-ISDN to B-ISDN

The N-ISDN technology has apparently assumed the characteristics of a Telecommunication Standard on a nearly universal scale, at least in the period of early nineties. The presentation of the N-ISDN as a universal network with one universal access to all services needs, however, to be modified. The basic concepts of the N-ISDN need vast enhancement to incorporate the requirements in the area of broadband services. Various proposals regarding broad-band solutions are presented and some are implemented (e.g. H-channels, frame relay, ...). The objection to a number of the methods is lack of universal solutions that are service-independent. (This is not so of course for all methods.)

It is, however, a paradox that demand for a complete new concept arises even before the present N-ISDN has matured and proved its merits.

Obviously the investments into the N-ISDN will greatly influence the introduction of ATM based B-ISDN network of the future. Backward compatibility and/or interworking need to be studied in detail to cater for existing and future customer demands.

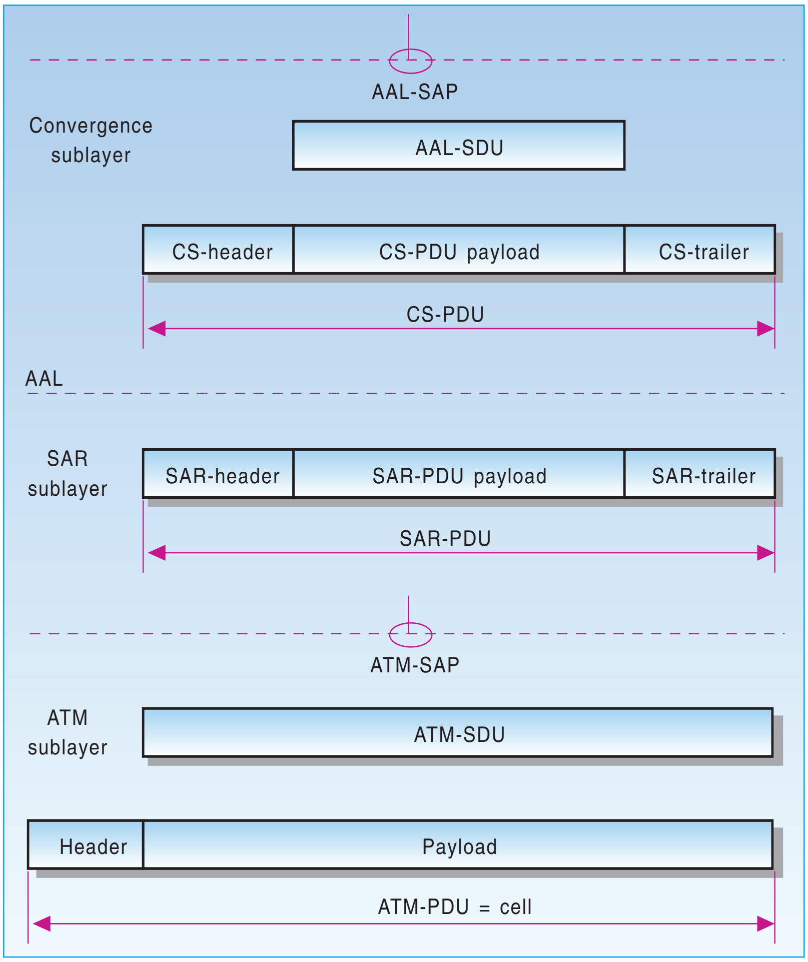


Figure 25 Data unit naming convention

The ATM solution rendering possible integration of transport for multirate traffic has gained recognition on an international level in organisations, major suppliers and operating companies. The CCITT is conscious of its role in the field of standardisations and a number of recommendations are issued concurrently with the development of further specifications for this new network generation providing the first truly services integrated network.

References

CCITT-Recommendations I.121, I.150, I.211, I.311, I.321, I.327, I.361, I.362, I.363, I.413, I.432.

Posisjonsberegning med tidsdifferansehyperbler

AV ÅGE ERIKSEN

1 Innledning

Dersom et problem ikke lar seg formulere eksakt matematisk, kan løsningen være en algoritme som benytter en repetert ledet søkesekvens i det aktuelle løsningsrommet. Slike metoder aktualiseres av tilgangen på rask regnekraft: Det spiller ofte ingen rolle om løsningen kommer etter 2 eller 20 000 iterasjoner.

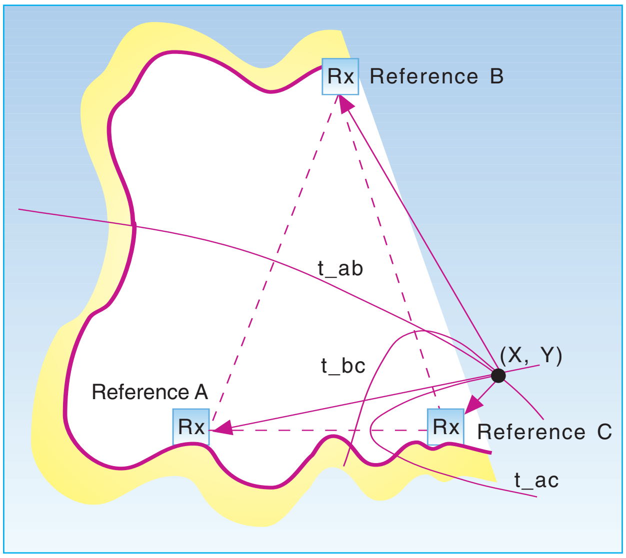
Som et eksempel på en slik innfallsvinkel er det i det følgende vist en algoritme som posisjonsbestemmer et objekt ved hjelp av tidsdifferansehyperbler. Algoritmen er generell og kan benyttes i et hvilket som helst system basert på tidsdifferansemålinger, uansett hvilke signaler og målemetoder som benyttes, og uansett hvor mange referansestasjoner som er involvert.

Algoritmen er benyttet i et testsystem for digital posisjonering, første gang i mars 1990 (se f eks B Forssell: *Bruk av Digital Sel-Call for automatisk posisjonsbestemmelse,* NORNA 88, Strömstad, 1988). Prosjektet er et samarbeid mellom Televerkets forskningsinstitutt og SINTEF-DELAB.

Algoritmen som beskrives i det følgende ble benyttet under utvikling og test av dette systemet. Tidsdifferanser og avstandsdifferanser er oppfattet som ekvivalente størrelser under forutsetningen av konstant lyshastighet.

2 Utgangspunkt

Følgende situasjon er gitt (figur 1): Objektet som skal posisjonsbestemmes sender ut en melding ved tidspunkt t_0. Denne mottas ved et antall referansestasjoner, benevnt *a, b,* og *c* ... , – eksempelvis kan dette være tre. Meldingene detekteres ved de absolutte tidspunktene t_a, t_b og t_c. Fordi vi ikke kjenner utsendelsestidspunktet t_0, har ikke størrelsene t_a, t_b og t_c noen direkte mening. Vi må derfor ta utgangspunkt i deres innbyrdes differanser:



Figur 1 Tidsdifferansehyperbler og referansestasjoner. Objektet som skal posisjonsbestemmes har koordinater (x,y)

$$t_{ab} = t_a - t_b$$
$$t_{ac} = t_a - t_c$$
$$t_{bc} = t_b - t_c$$

hvor parametrene til høyre er de vi faktisk måler.

Bemerkning 1. Geometrisk kan denne situasjonen illustreres med følgende fortolkning: Tidsdifferansen $t_{ab} = t_a - t_b$ definerer alle punkter hvor avsenderobjektets avstand til *referansestasjon a* minus avsenderobjektets avstand til *referansestasjon b* er lik med størrelsen t_{ab}. Disse punktene beskriver en hyperbel. Samme forhold vil gjelde for de andre stasjonene, og oppgaven blir å finne skjæringspunkter mellom hyperblene.

Bemerkning 2. I teorien kan det hevdes at det foreligger redundans, ettersom en av tidsdifferansene over kan uttrykkes ved hjelp av de to andre. Men dette er sant bare dersom tidsdifferansemålingene er støyfrie. Realistiske signaler utsettes for støy, – det er derfor ønskelig å benytte bidrag fra så mange tidsdifferansehyperbler som mulig.

3 Problem

Hyperblene kan ha en hvilken som helst krumming og orientering. Det har derfor vist seg vanskelig å uttrykke løsningen eksakt matematisk. Kompleksiteten øker dramatisk med antall referansestasjoner.

Det er også uhensiktsmessig å benytte tilnærmingsmetoder. Praktiske forsøk har vist at det er umulig å oppnå konvergensforhold som gjelder for et ubegrenset geometrisk areal. Det har dessuten vist seg meget vanskelig å definere *kriteriene* for konvergens.

En algoritme basert på *minste kvadraters metode* optimaliserer både step-lengde og -retning (B Forsell: *Radionavigation Systems,* Prentice Hall, 1991), men forutsetter at avstand mellom startpunkt og skjæringspunkt er liten i forhold til avstanden til referansestasjonene.

4 Idé

Ideen som er benyttet i denne algoritmen er basert på følgende tankegang: Det er vanskelig å konvertere tidsdifferansene f eks t_{ab}, t_{ac} og t_{ad} (benytter nå *fire* stasjoner) til et geometrisk punkt *(x,y)* som svarer til senderobjektets posisjon. Å gå den motsatte veien derimot, er trivielt. Ettersom vi kjenner referansestasjonenes koordinater, kan vi ved å anta en bestemt posisjon *(x,y)* beregne t_a, t_b, t_c og t_d, og dermed også deres innbyrdes differanser, t_{ab}, t_{ac} og t_{ad}.

Dette åpner for følgende muligheter:

- Hvis vi velger et vilkårlig punkt i planet, (x', y'), kan vi av dette generere de tilhørende tidsdifferanser t'_{ab}, t'_{ac} og t'_{ad} og sammenlikne disse med de målte størrelsene t_{ab}, t_{ac} og t_{ad}.
- Videre kan vi også avgjøre hvilket av to vilkårlige punkter (x',y') og (x'',y'') som er mest attraktivt i forhold til det ukjente punktet ved å sammenlikne deres tidsdifferanser med de målte differansene. Denne operasjonen kan utføres mange ganger og med så mange punkter vi ønsker.

5 Implementering

Implementeringen kan tenkes sammensatt av tre separate deler: en initialisering, en generator og en komparator. Initialiseringen definerer startkriterier for iterasjonen. Generatoren genererer løpende nye punkter i planet, og komparatoren aksepterer/forkaster genererte punkter med tanke på konvergens. Dette er realisert som følger:

5.1 Initialisering

5.1.1 Startpunkt

Før iterasjonen starter, må det velges et startpunkt. Punktets koordinator er irrelevante, men en enkel løsning er å velge midtpunktet mellom referansestasjonene, eller rett og slett benytte koordinatene til en av referansestasjonene.

5.1.2 Step-størrelse

Step-størrelsen s definerer den skrittlengden generatoren benytter til å generere nye punkter i planet. Step-størrelsen settes initielt til å være relativt stor, eksempelvis lik avstanden mellom to referansestasjoner. Generatoren justerer selv s til en hensiktsmessig størrelse under iterasjonen.

5.2 Generator

Ved første gjennomløp er n (index) lik 0 og punktet (x_0,y_0) er identisk med startpunktet.

Anta at generatoren ved iterasjon i opererer på punktet (x_n,y_n). Ved første påfølgende gjennomløp skal det genereres et sett nye punkter som grunnlag for videre iterasjon. Ettersom algoritmen skal ha full frihet til å søke i planet, kan følgende nye punkter genereres (figur 2):

- orientering mot nord: $(x_n,y_n + s)$
- orientering mot sør: $(x_n,y_n - s)$
- orientering mot øst: $(x_n + s,y_n)$
- orientering mot vest: $(x_n - s,y_n)$

hvor parameteren s er den valgte step-størrelsen.

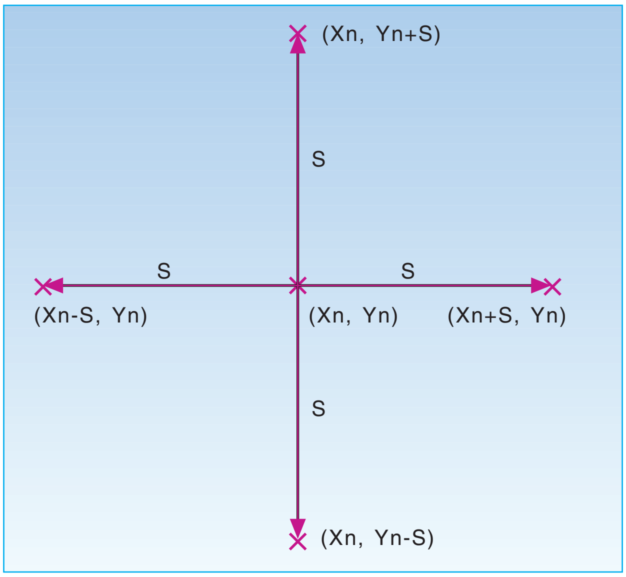
Komparatoren (omtalt nedenfor) utfører nå en test for å avgjøre hvilket av de totalt *fem* punktene som skal benyttes for videre iterasjon. Denne testen kan ha to mulige utfall:

Utfall I: Dersom et av de fire nye punktene blir foretrukket, eksempelvis kan dette være punktet som gir orientering mot vest, $(x_n - s,y_n)$, blir det nye punktet benevnt (x_{n+1},y_{n+1}), og generatoren vil ved neste gjennomløp generere et sett nye punkter:

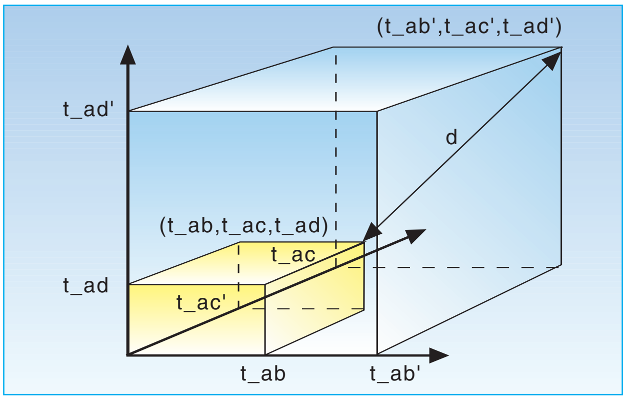
- orientering mot nord: $(x_{n+1},y_{n+1} + s)$
- orientering mot sør: $(x_{n+1},y_{n+1} - s)$
- orientering mot øst: $(x_{n+1} + s,y_{n+1})$
- orientering mot vest: $(x_{n+1} - s,y_{n+1})$

hvor parameteren s fremdeles er den samme step-størrelsen.

Utfall II: Dersom ingen av de fire nye punktene er å foretrekke framfor det foregående punktet (x_n,y_n), så trekker algoritmen



Figur 2 Punktet (Xn,Yn) med fire nye alternativer for videre iterasjon. Dersom ingen av de fire nye punktene aksepteres, så halveres s



Figur 3 Tidsdifferansestørrelser plottet i et koordinatsystem. Distansen d indikerer avstanden til hyperblene

følgende konklusjon: Step-størrelsen s er for stor og skyter over alle attraktive løsninger. Generatoren reduserer derfor step-størrelsen til det halve ($s \rightarrow 1/2\ s$) før et nytt sett punkter genereres som omtalt over.

Dersom reduksjonen av s fremdeles ikke leder til nye aktuelle punkter for videre iterasjon, foretas operasjonen ($s \rightarrow 1/2\ s$) på nytt det nødvendige antall ganger.

Når generatoren endrer s, får dette permanent virkning, dvs generatoren fortsetter med redusert step-størrelse i alle påfølgende iterasjoner. Etter p reduksjoner er step-størrelsen redusert til $1/2^p$ av sin opprinnelige verdi. Algoritmen terminerer når step-størrelsen er liten relativt til den nøyaktigheten som ønskes.

5.3 Komparator

Komparatoren avgjør hvilket av to punkter som gir kortest vei til løsningen. Dette gjøres på følgende vis:

I et vektorrom (figur 3) med akser som refererer til tidsdifferanser, f eks t_{ab}, t_{ac} og t_{ad}, kan de målte tidsdifferansene lokaliseres til et punkt med koordinatene (t_{ab},t_{ac},t_{ad}).

Et vilkårlig valgt punkt *(x',y') i planet* genererer tidsstørrelsene t'_a, t'_b, t'_c og t'_d som gir punktet $(t'_{ab}, t'_{ac}, t'_{ad})$ i vektorrommet. Euklidisk distanse mellom disse to punktene indikerer nå veien fra forslaget *(x',y')* til det ukjente skjæringspunktet *(x,y).*

$$d = \sqrt{(t_{ab} - t'_{ab})^2 + (t_{ac} - t'_{ac})^2 + (t_{bc} - t'_{bc})^2 + \ldots}$$

For hver iterasjon leverer generatoren totalt 5 punkter til komparatoren for test. Det punktet som gir kortest euklidisk distanse til løsningen i vektorrommet, blir foretrukket og returnert.

Bemerkning I: Når step-størrelsen er stor (relativt til hyperblenes krumming) vil det kunne skje at algoritmen periodevis synes å hoppe i en ugunstig retning i forhold til den ønskede posisjon. Dette kommer av at algoritmen søker kortest mulig distanse til *hyperblene,* ikke til løsningen. Først når step-størrelsen er redusert så mye at hyperblene i nærheten av sine skjæringspunkter fortoner seg som rette linjer i planet, vil det være sammenfall mellom distansen til hyperblene og distansen til målet.

Iterasjonens stoppkriterium innfris når distansen til hyperblene er minst, dvs i middelpunktet for hyperblenes skjæring.

Figur 4 viser et eksempel på hvordan algoritmen kan søke i planet fra startpunktet (x_0, y_0) og mot målet *(x,y).* Linjene nærmest *(x,y)* (når *s* er minst) er ikke tatt med pga oppløsningen i skissen.

Bemerkning 2: Det kan skje at punktet (x_n, y_n) og punktet (x_{n+1}, y_{n+1}) har den samme distanse til løsningen. Det vil da være en fare for at algoritmen skal kunne gå i vranglås ved å hoppe mellom punktene (x_{n+1}, y_{n+1}) og (x_n, y_n). Det er derfor nødvendig å innføre et krav om at komparatoren skal returnere et punkt bare dersom dette *definitivt* er bedre enn (ikke like godt som) sin forgjenger. Dette vil tvinge fram en reduksjon av step-størrelsen *s* før iterasjonen fortsetter.

6 Avsluttende bemerkninger

6.1 Prosesseringstid

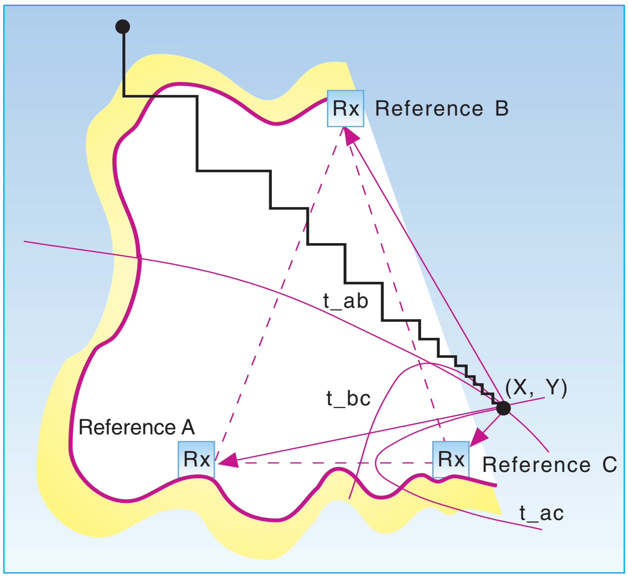
Som nevnt innledningsvis, benytter algoritmen en repetert ledet søkesekvens i det aktuelle løsningsrommet. Det spiller i prinsippet liten rolle om løsningen kommer etter 2 eller 20 000 iterasjoner. Praktiske forsøk (Oslofjorden) viser at algoritmen ofte terminerer etter ca 200 iterasjoner.

6.2 Multiple skjæringer

Under visse forhold vil det kunne foreligge situasjoner hvor hyperblene skjærer hverandre i flere punkter. Dette innebærer flere likeverdige løsninger, uansett hvilke metoder som benyttes.

I et praktisk system kan problemet håndteres på flere måter:

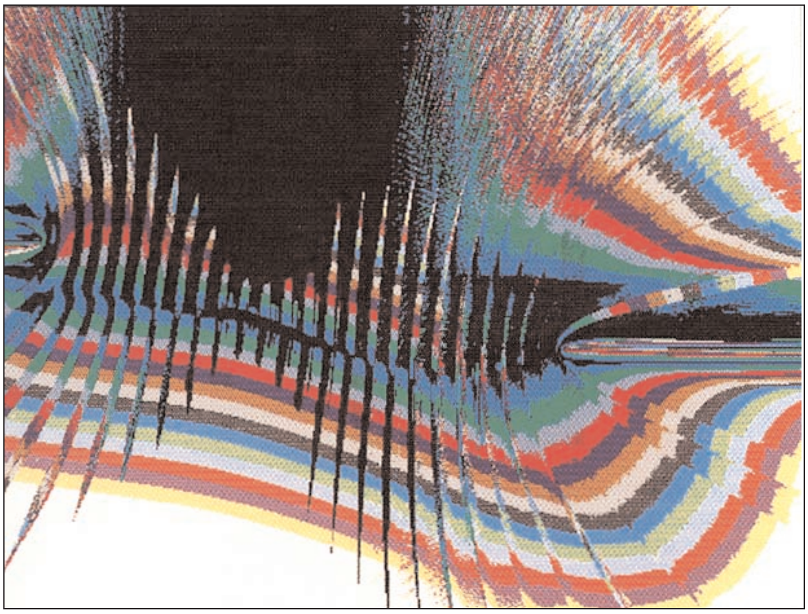
- Operatøren forsynes med et grafisk grensesnitt som markerer hyperblenes beliggenheter. Basert på dette kan algoritmen gis et startpunkt som svarer til den løsningen operatøren ønsker å bestemme.
- En algoritme kan bygges slik at den rekursivt finner alle likeverdige løsninger. Det vil likevel være nødvendig med inngrep fra en operatør for å bestemme hvilken av de foreliggende løsninger som skal benyttes.



Figur 4 Eksempel på en mulig søkevei fra startpunktet (Xo,Yo) mot målet (X,Y)

6.3 Bruk av flere referansestasjoner

Ettersom algoritmen under forhold med støy itererer mot *middelpunktet* av flere skjæringer, egner den seg godt i situasjoner med flere enn tre referansestasjoner. Det er i prinsippet ingen øvre grense for hvor mange referansestasjoner som kan benyttes. Bruk av mange referansestasjoner er enkelt å implementere.


Figur 5 Posisjonering i Oslofjorden med en oppløsning på 200 meter. Mørk farge angir liten iterasjonstid, lys farge angir lang, - størrelsesorden 200 iterasjoner. Metoden ble brukt til å kartlegge algoritmens konvergens-kriterier og definere falske løsninger

MMIC - new technology offers new opportunities

BY LEIF HANSSEN AND CHRISTIAN F HEIDE

621.38.049.7

1 Introduction

The development of integrated circuits has had a major impact on our every-day lives. Especially in telecommunications by supplying more and more complex services to an ever decreasing price.

The number of transistors on a chip was previously used as a figure of complexity, and the acronym VLSI was used for the most complex chips with thousands of transistors. Today it is more common to divide the circuits into two categories, standard circuits and ASICs. Number of transistors is no longer considered as a figure of complexity. In microwave applications this is obvious since the number of transistors usually is very small while the complexity can be quite high.

1.1 The concept

MMICs are integrated circuits designed to operate in the GHz frequency range. The term "monolithic" indicates that the whole circuit is integrated on a single semiconductor chip. The components are not soldered to the chip, but formed using the semiconductor actively. This is contrary to MICs where standard components such as transistors, capacitors and resistors, are soldered to a microwave substrate. In MICs components do not have to be planar which is the case for MMICs.

1.2 History

The idea of monolithic integration of microwave circuits was given birth in 1964. A research programme funded by the US government had as its objective to realise a transmit/receive module for an aircraft phased-array antenna (1). The researchers of the programme tried to build this module by using monolithic integration of components on a semi-insulating silicon substrate. The results were, however, not very encouraging, mainly because the semi-insulating properties of silicon were not good enough. Because of this and other problems, the work on MMICs lay dormant until 1967-68 when Mehal and Wacker tried to use gallium arsenide as substrate material with Schottky and Gunn diodes as the active components (2). This was far more successful because of the good semi-insulating properties of gallium arsenide. However, it was not until 1976, when research workers at Plessey applied this approach to an X-band amplifier based on the MESFET as the active component, that the present intense activity began (3).

Today MMICs are made mainly by using gallium arsenide as substrate material and MESFETs as the active components. There is, however, a growing interest in other transistor structures, especially HEMT and HBT.

1.3 Why monolithic?

Monolithic circuits have several advantages as compared to hybrid ones. One great advantage is the possibility of processing a large number of identical circuits in parallel, which will in turn lower the price per circuit. Another advantage is that the number of interconnections and bonding wires required is greatly reduced. Furthermore, one does not have to do time-consuming soldering and tuning. MMICs are generally very small and light-weight, and they can exhibit high reliability and high reproducibility. A relatively large circuit-design flexibility is obtained and there are possibilities for multifunction performance on a chip. In the future it will be common to have a complete transmit/receive-module on a single chip, or, alternatively, as a multichip MMIC module (4).

The production cost for a circuit is given by the size of the circuit and the yield, which means that it is important to get many functions on a small area. Since the number of functions does not determine the cost, moving the border of integration by integrating more functions on the same chip can give extra functions for free. There will be no cost for the extra functions added to the chip except for the increase in chip area and a possible decrease of yield. Another reason for monolithic integration is the miniaturisation, to get a smaller size, small volume/area and low power consumption. This can improve the performance of the system since it decreases the length of the signal paths. Interconnection between different chips do degrade the signal. This means that a monolithic circuit with many integrated functions can have a better performance than a discrete system built with functions where each of these functions have a better performance.

MMIC has one advantage compared to MIC which is rarely considered. This advantage which it has in common with, and which it has adopted from other ASIC-designs, is the fact that the specification of the system is developed at the simulator. In traditional discrete electronics the system was often developed at the test bench and the specification was finished when the prototype system was built and characterised. This procedure cannot be used in ASIC design since the fabrication of circuits takes a fairly long time and is quite expensive.

The drawbacks associated with the monolithic approach are mainly the difficulties with trouble-shooting and tuning. MMICs generally have low Q-factors. This implies that they will exhibit good broad-band performance, but make poor resonators.

2 Transistors

2.1 Properties of gallium arsenide

Gallium arsenide (GaAs) is a so-called III-V semiconductor and is ordered in a zincblende lattice. This means that the material is made using elements from main group III (gallium) and V (arsenic) in the periodic table of the elements. The resistivity of undoped gallium arsenide is in the order of 10^8 Ωcm, which is quite high compared to silicon (about 10^4 Ωcm) and other semiconductors (5). Pure GaAs is therefore often referred to as semi-insulating.

A simplified energy band diagram of GaAs is shown in figure 2.1 (6). The conduction band has three valleys, named Γ, L, and X. The Γ valley has the lowest minimum of the three, and so determines the bandgap which is 1.42 eV as compared to 1.11 eV for Si. The larger bandgap implies that GaAs is more resistant to radiation and may be operated at higher temperatures than Si. It is evident from figure 2.1 that the bandgap is direct, i.e. a transition from the valence band to the Γ valley requires

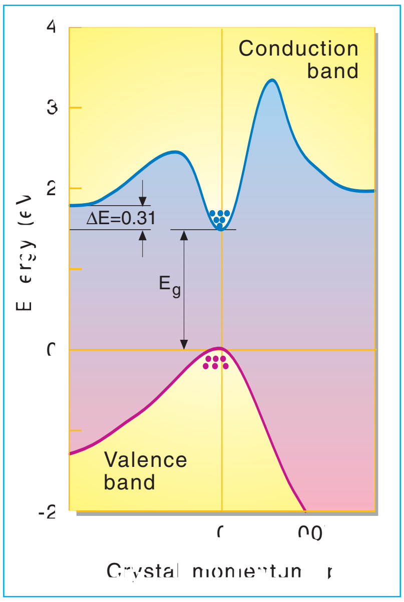


Figure 2.1 Simplified energy band diagram of GaAs

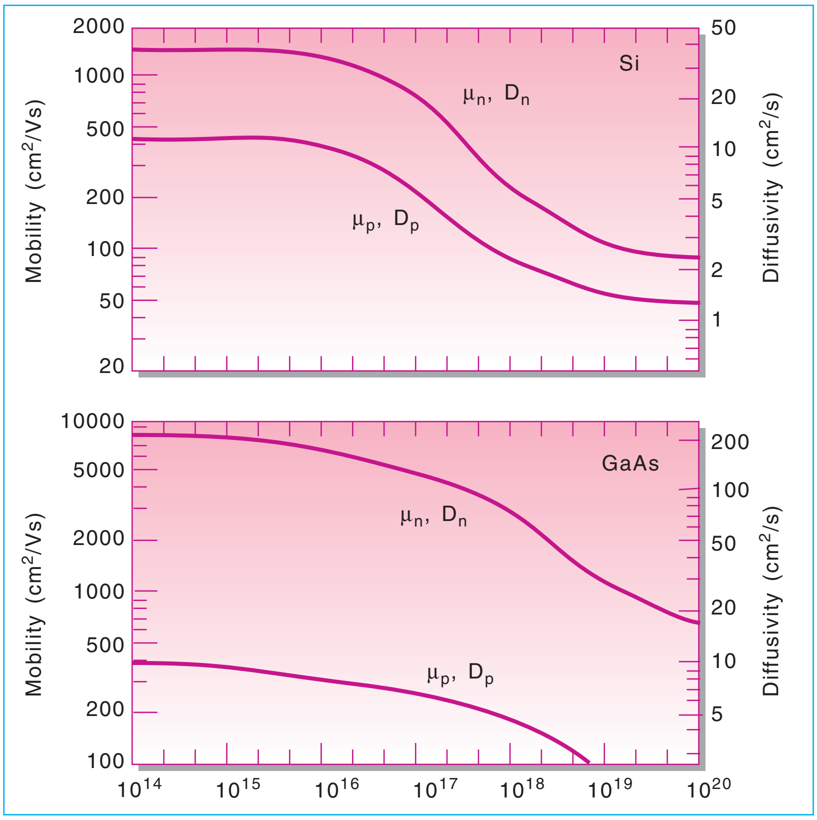


Figure 2.2 Mobility and diffusivity of GaAs and Si as a function of the doping level

no change in crystal momentum. This makes possible transitions upon exchange of photons. The bandgap of GaAs equals a wavelength of 875 nm. The electrons in the Γ valley have the smallest effective mass since the effective mass is smaller for the narrower valley (5). The effective mass of electrons in the Γ valley is 0.07 m_0, where m_0 is the free electron rest mass (7). This yields a very high electron mobility of GaAs, namely 8500 cm^2/Vs (5). The effective mass of the holes are 0.5 m_0 and the hole mobility is around 400 cm^2/Vs. For silicon the mobility is 1350 cm^2/Vs and 480 cm^2/Vs for electrons and holes, respectively. These figures explain why only electrons are used as carriers in GaAs. This also prohibits a complementary technology, like CMOS, to be used in GaAs.

The mobility is reduced when the semiconductor is doped. For heavy doping the electron mobility is reduced to about 2200 cm^2/Vs (7). Figure 2.2 (5) compares the mobilities and diffusivities of GaAs and Si as a function of the doping.

The most frequently used dopant for GaAs is silicon. The ionisation energy for this donor is approximately $6 \cdot 10^{-3}$ eV (5), which is so small that virtually all donors will be ionised at room temperature.

If some of the Ga atoms are replaced by aluminium (Al) atoms, the bottom of the valleys are elevated, resulting in a larger bandgap. This is utilised in so-called heterostructures, where materials with different bandgaps are put together to achieve certain advantages. Examples of components with heterostructures are HEMTs and HBTs. It is important to note that the Γ valley rises more slowly with aluminium concentration than do the two others (6). This is illustrated in figure 2.3. The result of this is that the bandgap becomes indirect for x > 0.45, where the X valley becomes the lowest conduction band valley.

The heterostructures which the HEMTs are based upon, are fabricated using MBE or MOCVD. At our MBE-lab, material with

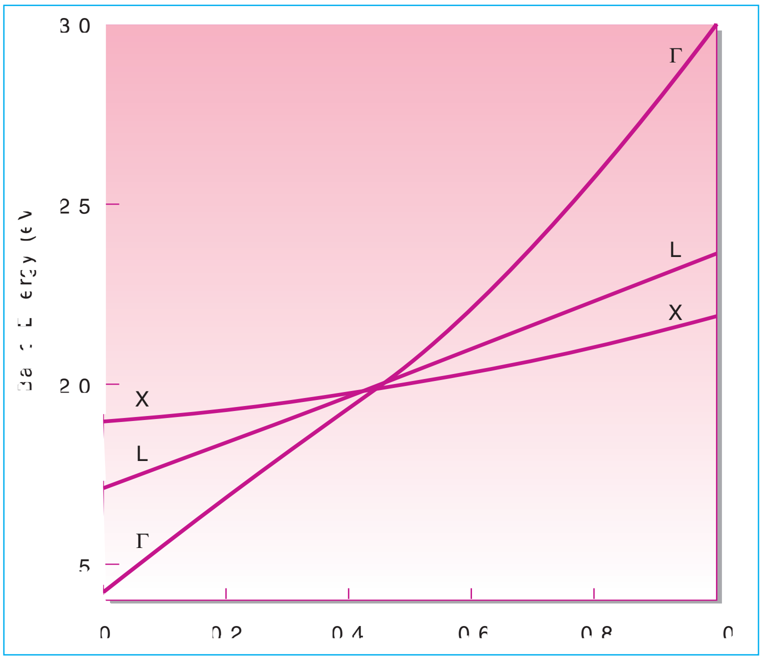


Figure 2.3 Variation in Γ, L, and X valley energies with aluminium band fraction, x, in $Al_xGa_{1-x}As$

electron mobilities as high as 283,000 cm^2/Vs at 77 K has been grown (8).

GaAs is well suited as a microstrip substrate as well. High resistivity of the substrate implies small transmission losses. The dielectric constant is relatively high (around 13). This helps in

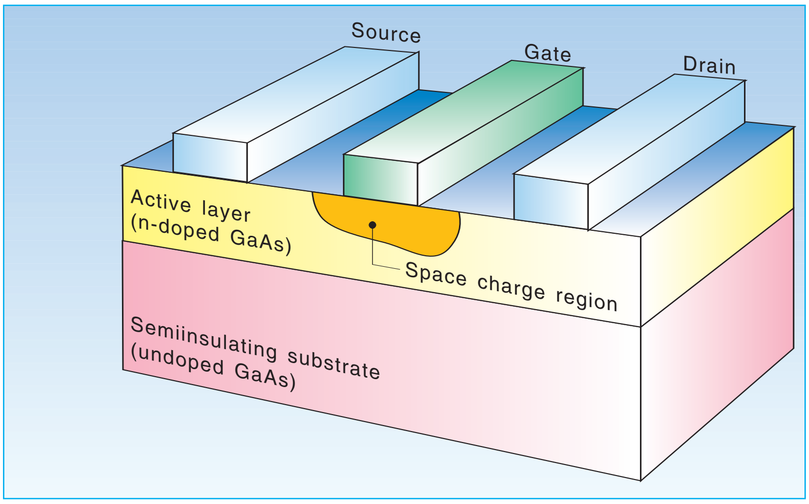


Figure 2.4 Cross section of a MESFET

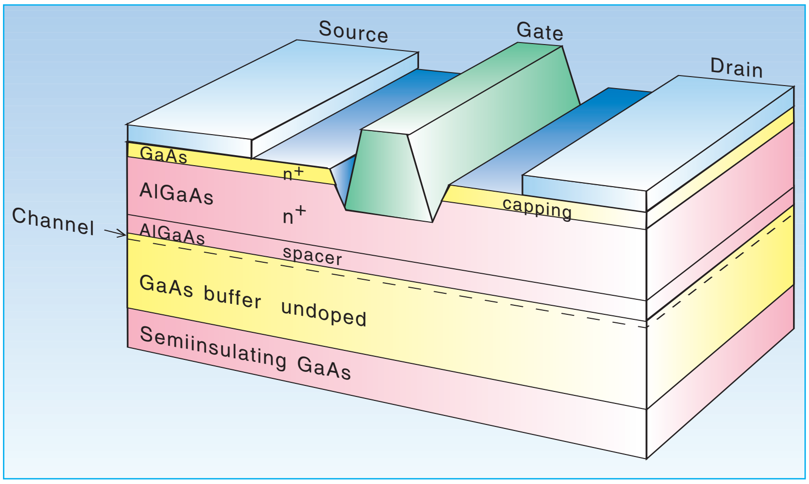


Figure 2.5 Cross section of a HEMT

miniaturising the circuits since the wavelength in a material is inverse proportional to the square root of the dielectric constant. One disadvantage of GaAs is a very high thermal resistance which can cause heat problems for high power amplifiers.

2.2 MESFET

The MESFET is the most widely used transistor in MMICs. The MESFET is made, as indicated in figure 2.4, by placing a doped GaAs layer on an undoped layer of the same material. Since the resistivity of undoped GaAs is so high it effectively isolates the MESFET from the surroundings. As the name indicates, the gate electrode of the MESFET is placed directly on the semiconductor surface (figure 2.4) in contrast to, for instance, MOSFET transistors. This is possible because the metal/semiconductor junction gives a Schottky diode which conducts literally no current when reverse biased. The depletion region that appears at the Schottky-junction decreases the effective channel cross section. Any voltage applied at the gate modulates this depletion region. When a DC-voltage is applied at the drain electrode, the current flowing through the channel is proportional to the gate voltage. The source and drain electrode/semiconductor junction is made ohmic. The doping of the active layer is always of n-type because of the low hole mobility in GaAs.

The high-frequency performance of MESFETs is mainly determined by the gate-length: Shorter gates give higher cut-off frequencies. Transistors commonly offered by GaAs vendors, have gate-lengths of 0.5 µm, and yield an f_t in the order of 20-25 GHz and a transconductance in the order of 250 mS/mm.

2.3 HEMT

HEMTs (also called MODFET, TEGFET, SDHT and HFET) are a kind a field-effect transistors with improved frequency and noise characteristics as compared to MESFETs. The enhanced electron mobilities in modulation doped heterostructures and superlattices, were first observed by Dingle et al in 1978 (9). The basic idea of the HEMT is to separate the charge carriers from the doping impurities. This is favourable since the donors act as impurities and scatter the electrons resulting in loss of electron velocity. Figure 2.5 shows a cross section of a HEMT. On top, there is a highly doped GaAs layer. This helps in getting good ohmic contact to the electron gas. Underneath there is a region of AlGaAs highly doped with Si. Si acts as a donor in AlGaAs. Then there is a thin layer of undoped AlGaAs, called the spacer layer. This further separates the charge carriers from the donors. Underneath there is a layer of undoped GaAs. Because AlGaAs has a larger bandgap than GaAs, the electrons near the junction are trapped in an extremely thin layer (potential well) close to the junction in the undoped GaAs. This can be seen from the band diagram of the junction, shown in figure 2.6. The width of this well is in the order of 80 Å. Because of this, the charge carriers are not subject to impurity scattering from the donors, and the carrier transport is mainly limited by lattice scattering. Because of this, and because the dimensions of the transistors decrease, the mean free path may become greater than the gate-length. Hence, the transport mechanism may be quasi-ballistic resulting in a large velocity overshoot and increased frequency response.

In the mid-1980s it was recognised that the improved transport properties in a HEMT do not necessarily result in improved performance as compared to the MESFET (10). The components' ability to modulate the high-speed electrons must be taken into account as well (11). This is the main reason why so-called pseudomorphic HEMTs based on AlGaAs/InGaAs/GaAs, have better performance than the AlGaAs/GaAs HEMTs (10), (12). The term pseudomorphic indicates that there is some lattice mismatch between the substrate (GaAs) and the epitaxial layer (InGaAs). The channel of a p-HEMT provides enhanced mobility (1.3 times greater) and electron velocity (1.5 times greater) compared to a conventional HEMT (12).

Recently, there has been fabricated lattice matched HEMTs with InAlAs/InGaAs grown on InP substrates instead of GaAs (13). This type of device offers higher electron sheet charge density and better carrier confinement in the channel (13). This results in superior electron transport properties which yield higher transconductance, higher cut-off frequency and lower noise than GaAs based p-HEMTs.

The highest extrinsic f_t reported to date is 340 GHz (14) which equals an intrinsic f_t of 550 GHz. This device has a transconductance of 1740 mS/mm, and a gate length of 50 nm.

In figure 2.7 the minimum noise figure (NF_{min}) and the maximum available gain (G_{ma}) of a typical MESFET, a typical HEMT and a typical pseudomorphic HEMT is compared (15). The gate length is 0.25 mm for all three of them. As can be seen, the p-HEMT is best in both respects. It has a cut-off frequency *(f_t)* of 230 GHz.

Until now it has been known that HEMTs have better noise properties than MESFETs. However, a recently published letter (16) has reported a MESFET having noise properties comparable to the best HEMTs. This MESFET had a gate-length of 0.25 µm and is fabricated using ion implantation. The noise figure was 0.56 dB at 10 GHz with 13.1 dB associated gain. At 17 GHz the noise figure is 0.83 dB with 10.5 dB associated gain. The extrapolated cut-off frequency was in the order of 80 GHz (16). The major advantage of this MESFET compared to HEMTs, is the cost, manufacturability, and robustness of the fabrication process (16). The authors of (16) have also made an equivalent MESFET with a gate-length of 0.15 µm (17) with 0.6 dB noise figure with an associated gain of 17 dB at 10 GHz. The extrinsic cut-off frequency was 109 GHz. These measurements indicate that the average electron velocity under the gate is determined primarily by the high-field electron velocity rather than the low-field mobility (18).

2.4 HBT

As the name indicates, the HBT is a bipolar transistor, contrary to the MESFET and the HEMT (19). It is most commonly fabricated using AlGaAs/GaAs. Figure 2.8 shows a cross section of an HBT. The heterojunction is formed at the base-emitter interface. The holes will then see a much higher potential barrier at the junction, which allows for a much higher p-doping of the base since the leakage of holes from the base to the emitter is reduced. This leakage current is a severe limitation in conventional npn-transistors. This higher p-doping makes it possible to reduce the base-width since the depletion-regions at the junction is reduced when the doping is increased. The base may be as small as a few hundred Ångstrøm. The smaller base improves the high frequency performance of the device. This is achieved with conventional optical lithography.

In general, HBTs have higher transconductance (10 to 100 times) than MESFETs and HEMTs (19). This permits smaller input voltage swing and yields low output impedance in a common-collector configuration. In a common-emitter configuration, however, the output impedance is higher than that of MESFETs and HEMTs (19). This yields high device linearity and high DC voltage gain. Since the threshold voltage of an HBT is determined by the bandgap (19), it has a much more uniform distribution across a wafer as well as from wafer to wafer. Furthermore, the flicker noise (1/*f* noise) is lower in HBTs as compared to MESFETs and HEMTs. HBTs also allow higher current per effective transistor area, and thus a higher power handling capability (19). They also have greater radiation hardness than MESFETs and HEMTs, which is important in space applications.

However, the HBTs are not particularly suited for low white-noise applications. There are also some problems associated with the processing of such devices.

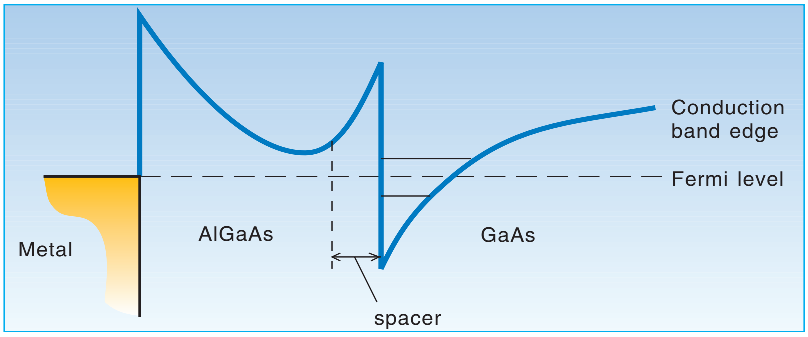


Figure 2.6 Band diagram of a HEMT in thermal equilibrium

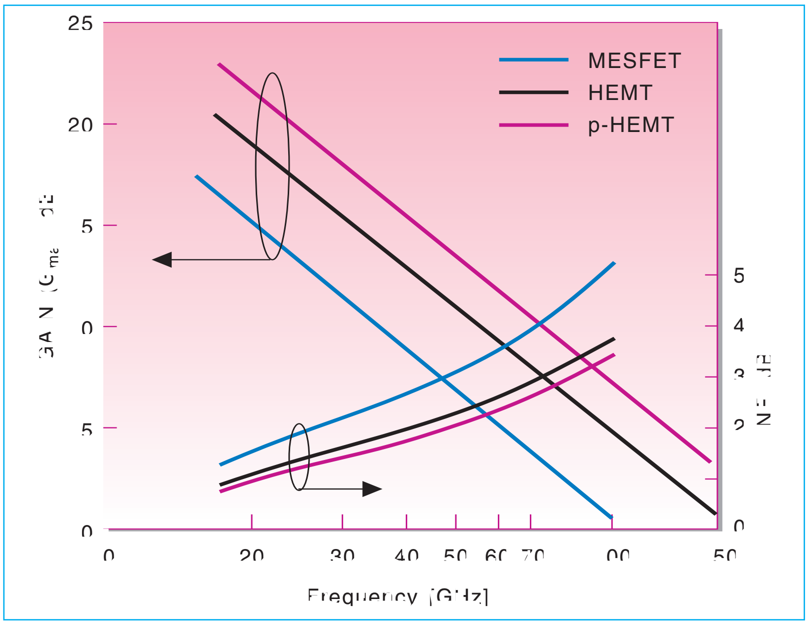


Figure 2.7 Comparison of a MESFET, a HEMT, and a p-HEMT, all with a gate-length of 0.25 µm

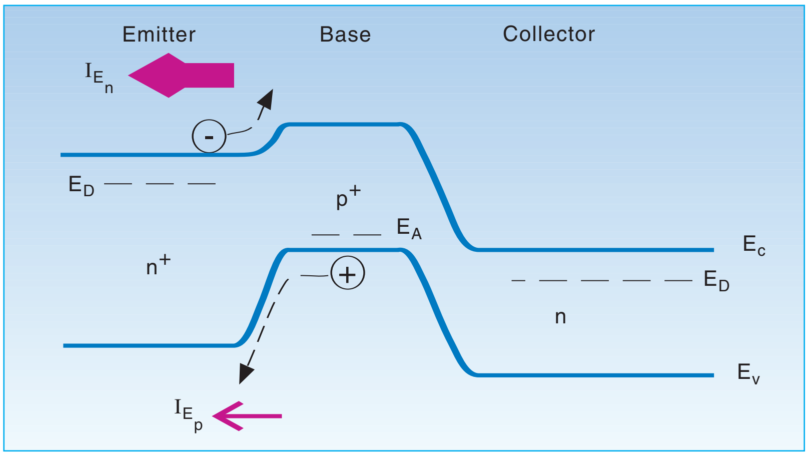


Figure 2.8 Cross section (a) and band diagram (b) of an HBT

3 Passive components

Availability of good passive components is a necessity to make high quality MMICs. These components are required to bias the active elements and to modify their characteristics by use of feedback. Reactive components are used to make the feedback frequency selective. The active components have a complex input and output impedance and passive components are

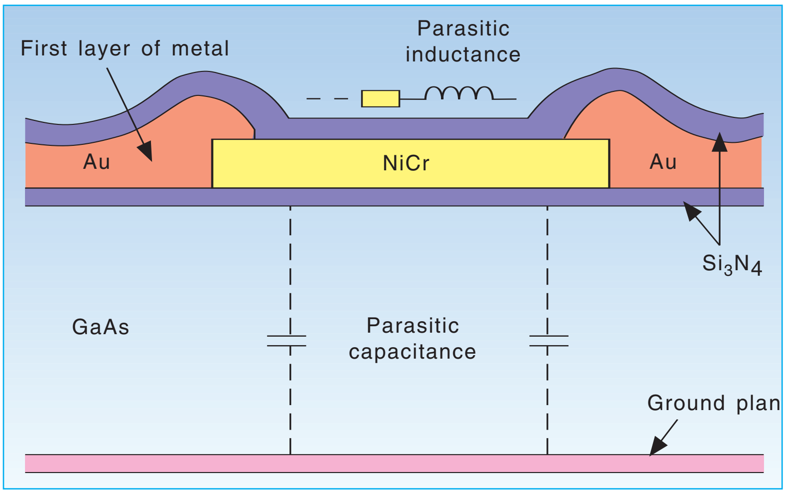


Figure 3.1 Implementation of a NiCr resistor

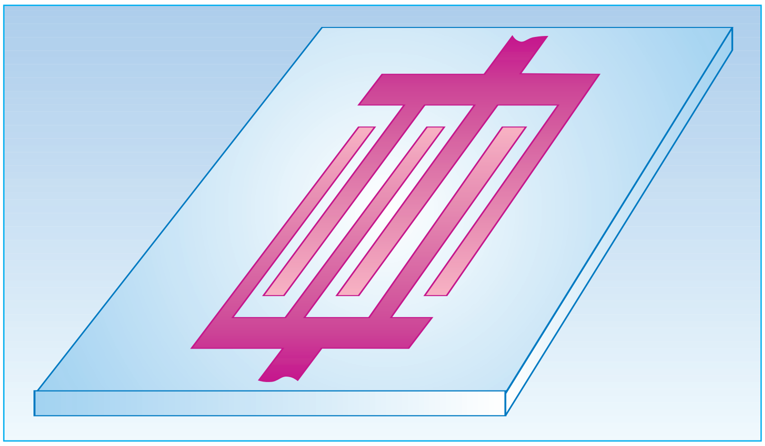


Figure 3.2 An interdigitated capacitor

required to make interstage matching network and to match the input and output of the circuits to 50 Ω.

Transistors are mainly used as gain elements. To give these elements optimal working conditions it is necessary to have access to passive components of high quality. Furthermore, the properties and accuracy of the passive components often limits the performance of the whole chip.

In MMICs there are usually four types of passive components: resistors, capacitors, inductors, and transmission lines. The way they are implemented on the chip differs from vendor to vendor and depends on the applications for which the process is intended.

3.1 Resistors

Resistors are mainly used to bias transistors. They can also be used to add loss to a circuit in order to stabilise it. However, as a general rule, resistive loss is undesired since it increases the noise in the system.

One common way of making good resistors is by adding one extra mask step to the process by deposition of a NiCr layer. NiCr is a well known material from thin film technology and has a well defined electrical behaviour. Typical sheet resistance can be in the order 50-200 Ω/square with an absolute accuracy of 10 % and a relative accuracy better than 1 %. In this context the notion absolute accuracy means the tolerance of the component while relative accuracy means the match of two resistors in close proximity of each other. Good matching requires wide devices (>10 µm). By use of laser trimming, L-cut method, the accuracy can be as good as 0.01 %. Of course, the disadvantage of laser trimming is very high production costs. Other advantages of NiCr is low voltage and temperature coefficients. One parameter which is important in design of integrated circuits is current density. This will define the minimum width of the resistor.

For small resistors a lumped resistor element is adequate as a simulation model. For large devices a more complicated model must be used at high frequencies. The parasitics of a NiCr resistor which must be taken into account are capacitance to the ground plane and series inductance in the conductor. As an example consider a minimum width resistor of 5 kΩ. A typical series inductance of such a device is in the order of 0.1 nH. This means that in most practical cases a simple lumped resistor model will do. Special attention must be made only for very long resistors. When the resistor is longer than one tenth of the wavelength then a distributed model, such as a lossy transmission model should be used. Such a model exists in most linear MMIC simulators.

Tantalum is another resistive material which is rarely seen in this context. Tantalum has a wider range of sheet resistance and an accuracy comparable to that of NiCr. Trimming can be done with oxidation or annealing by local heating with a laser.

In most processes the implant layers may be used for resistors. The highly doped N+ implant has a sheet resistance similar to that of NiCr and is generally not used when NiCr is available. The lightly doped N- implant layer has a sheet resistance which is typically one decade higher than that of NiCr, but with a much poorer accuracy. This layer is also susceptible to backgating effects and its use as a resistor is in practice limited to non-precision DC resistors.

3.2 Capacitors

Capacitors are required in MMICs for DC-blocking between different stages and at the input and output of the circuit. They are also used for on-chip decoupling of the power supplies. In feedback loops and interstage matching network, capacitors are required for doing impedance manipulation and filtering.

In GaAs MMICs there are mainly two ways of realising capacitors. These are interdigitated capacitors which can be done in one layer, and overlay or MIM capacitors which require two interconnection layers and a good insulator.

Interdigitated capacitors, figure 3.2, are mainly used when there are only one layer of interconnection available, and for small value capacitors where a very high precision is essential. This capacitor cannot be simulated with a simple lumped capacitor but requires a more complex model. Such a model exists in most linear microwave simulators, but there are heavy constraints on the shape of the capacitor and the range of usage. The capacitance results from fringing between the interleaved fingers. Consider, as an example (20), an interdigitated capacitor with 58 fingers of 60 µm, a finger width and finger space of 2 µm. The metal thickness is 0.1 µm. This gives a capacitance of 0.21 pF, with a Q factor of 100, a series inductance of 23 pH and a capacitance to the ground plane of 14 fF. The capacitor has a series resonance of 72 GHz and covers an area of 0.02 mm^2. The latter figure indicates the disadvantage of the capacitor: It is very area consuming. These figures equal an artificial sheet capacitance of 10 aF/μm^2.

Overlay or MIM capacitors, figure 3.3, are preferable in most applications because of their high area efficiency. GaAs does not form a natural oxide that can be used as an insulator which is the case for silicon (SiO_2). This means that the insulator cannot be grown but must be deposited. There is a practical lower limit for the thickness of such a layer to ensure a high yield for large capacitors. A frequently used insulator, is silicon nitride (Si_3N_4) with a thickness of 200 nm. This gives a typical sheet capacitance in the order of 300 aF/μm^2, with a yield better than 75-90 % for capacitors as big as 50 pF. In the MMIC context 50 pF is a very large capacitor. For low frequencies and for small size capacitors a simple lumped capacitor can be used in simulation. For larger capacitors and/or higher frequencies, the parasitics must be taken into account. These include the capacitance between the lower capacitor plate and the ground plane (C_{pl}), the fringe capacitance between the upper capacitor plate and the ground plane (C_{pu}) and the series inductance in the capacitor plates. As an example, a 50 µm · 50 µm MIM-capacitor on a 100 µm thick GaAs substrate has a capacitance of about 0.7 pF. The series inductance is 14 pH which gives a resonance frequency of 35 GHz.

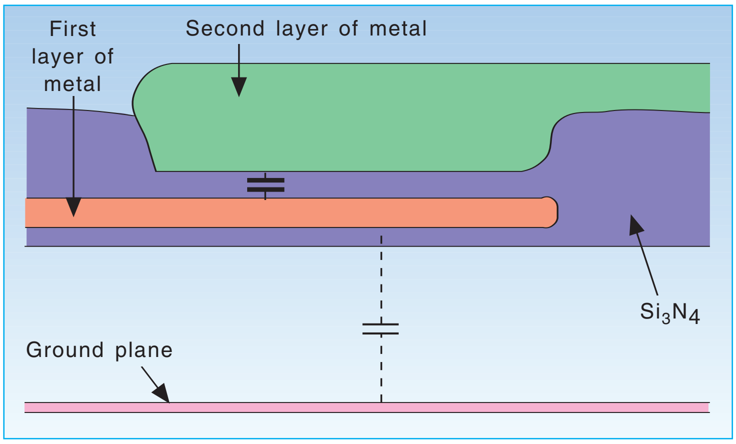


Figure 3.3 An overlay capacitor

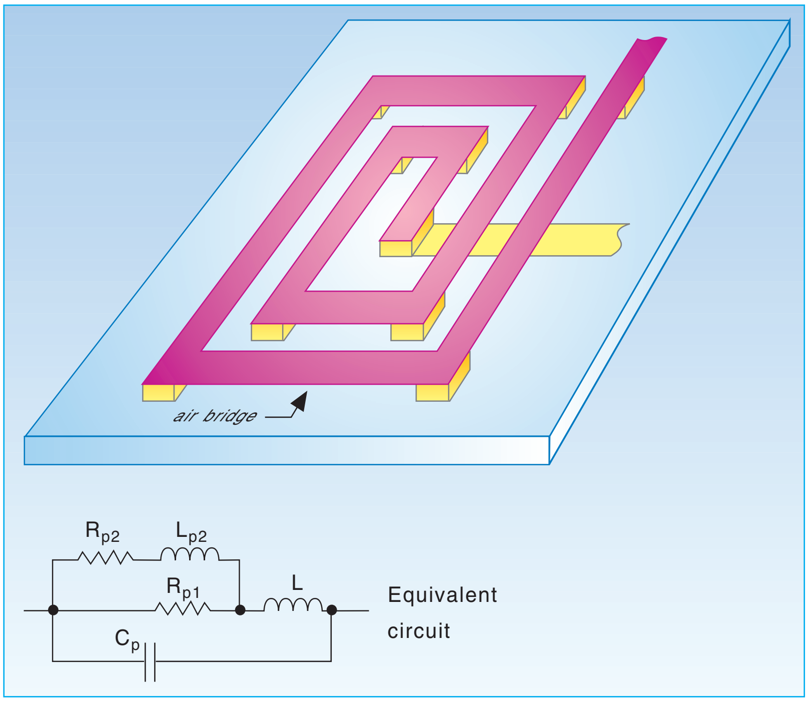


Figure 3.4 Spiral inductor: a) Layout b) Equivalent circuit

3.3 Inductors

There are several different ways of realising planar inductors in monolithic circuits. An inductor is based on mutual coupling between various line segments of a metallisation layer. The shape of the inductors depends on how many layers of metallisation that are available. With only one layer of metallisation the meander line and S-line is the most common realisation. Today, most processes offer two layers of metallisation which makes it easy to implement spiral inductors (figure 3.4). The extra layer of metallisation is required to contact the inner terminal. The inductor can have a circular or rectangular shape, where the latter is the most commonly used. If the second metal layer is an air-bridge metal the parasitic capacitance to ground is reduced and hence the resonance frequency is increased. Such an inductor is standing on poles of the first metal layer.

A lumped inductor model does not provide the required accuracy for most kind of simulations. The exception is for very small value inductors at low frequencies. The parasitics which must be taken into account are the series resistance and the capacitance from the input to the output. A possible model for simulation is shown in figure 3.4b. From a library containing a discrete number of inductors the figures for the chosen inductor can be found in a table and used in simulation. The correctness of the model can be verified against S-parameters extracted by measuring an inductor realised on a chip.

Typical inductor values for MMICs range from 0.1 to 10 nH. High value inductors are usually realised by a thin metal line. For these elements the current density must be considered. Accuracy, both absolute and relative, is very good. Any variation must be in metal width and thickness which mainly influence the resistance. This means that such elements have a very high reproducibility.

3.4 Transmission lines

At microwave frequencies, any conductor have both series inductance and parallel capacitance to ground. The capacitance is controlled by deposition of a metallic ground plane onto the backside of the chip. This means that it must be characterised as a transmission line. This form of transmission line is known as microstrip, and has been well characterised for MICs. In a two-

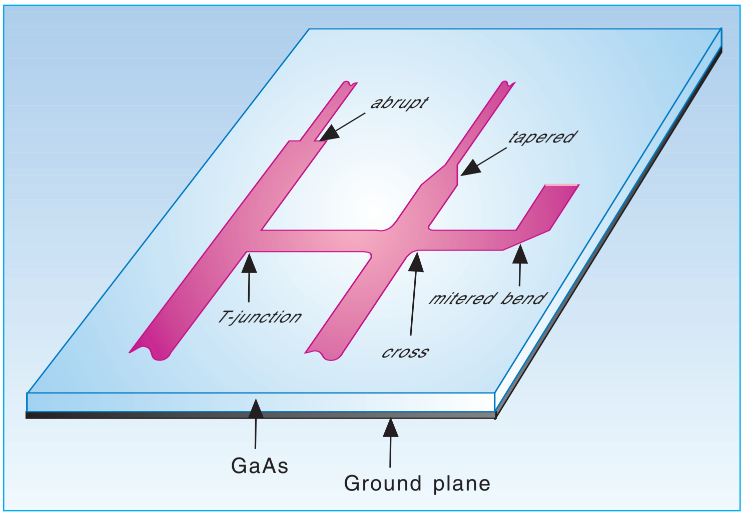


Figure 3.5 Different microstrip line shapes

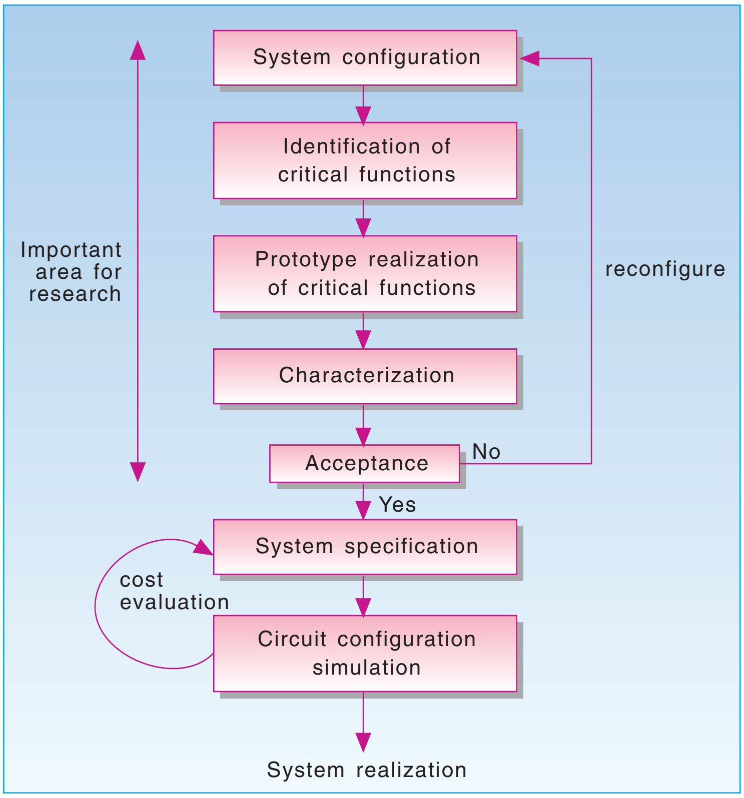


Figure 4.1 A part of the system development

layer metallisation MMIC technology there are mainly two configurations of transmission lines. The most straightforward realisation is by using the first interconnection layer. The other method is to use a so-called landed airbridge. This lowers the series resistance and inductance by increasing the thickness of the conductor and hence gives a transmission line with a different electrical behaviour. The airbridge metal can be handled as a wire if special care is taken. Considered as a transmission line, the low parasitic capacitance to ground reduces the transmission line effect. These facts mean that if the length of the airbridge is a small fraction of the wavelength, it can be modelled as a wire.

At high frequencies, the shape of the microstrip line has considerable effect on the response of the line. All kinds of shapes are described mathematically such as bends and corners, both mitered and curved (see figure 3.5). This is also the case for connections such as abrupt, tapered, T-junction and crosses. Models for all these microstrip elements are included in most simulators.

3.5 Other components

In microwave applications there are a lot of special components such as tunnel diodes, IMPATT-diodes, circulators and isolators, resonators, varactors, baluns, and so on. Most of these components cannot be integrated on a monolithic circuit. The exception here is Schottky diode which also can be used for varactors. A Schottky junction is required for making the MESFET gate and comes for free in a GaAs MESFET process.

4 Design of MMICs

The design of monolithic integrated circuits for microwave applications is somewhat different from design of circuits working in the kHz and MHz range. This is due to a different electrical description of components at these frequencies. All kinds of parasitics must be taken into account, which means that most components will have a non-ideal behaviour. These factors are something that MMIC designers share with MIC designers, which means that the MMIC designer can use the same kind of simulation tools as the MIC designer. However, there is one major difference. When the MIC is designed, it can be tuned during the test to make it meet the specifications. Very often the MIC designer takes this into consideration when designing the circuit. In monolithic circuits tuning is greatly undesired since it is difficult to do in practice. Secondly it increases the production costs considerably and removes one of the most important reasons for choosing a monolithic solution. As a rule of thumb, tuning of monolithic circuits should be avoided.

Since it is difficult to tune the circuit after fabrication and since the fabrication is very expensive and takes a long time, a trial and error procedure with several rounds of prototyping is out of the question when designing MMICs. This means that most of the trials must be done in the simulator and that errors must be minimised prior to fabrication. For a successful design it is an assumption that the designer is familiar with the simulation tools and knows their possibilities and limitations.

When designing MMICs, the choice of circuit configuration and the way the problem is solved, depends on the fabrication process. The components available and the performance of these components also determine the performance that can be expected from the circuit. The limited number and the poor accuracy of the components, compared to discrete components, limit the performance of the circuit.

Usually there are three possible design levels when working towards a foundry. The first level is that the customer only does the specification and then the foundry does the rest. That is, the design, the fabrication, the packaging and the test. This is the easiest way for the customer if he does not have experienced designers in-house. The second level is that the customer does the design, but with support from the foundry. The foundry does the fabrication and the customer does the packaging and the test. The third level is that the customer does everything except the fabrication. What level to choose depends on the design-experience of the customer. A very experienced designer who knows the process very well will take a high responsibility for the circuit design.

There is one major difference in the microwave simulation tools such as EEsof's Libra, as compared to analogue simulators for silicon circuit design such as SPICE. This is the possibility of doing design optimisation. The designer can choose a circuit configuration, specify the range of the components and the desired circuit performance. The simulator will then try to find component values that yield the specified performance. This procedure is not quite automatic and requires some experience from the designer so as to give realistic specifications. The optimisation is an efficient tool but it is still the designer who has to do the job.

The fact that the tolerances of the components are quite poor must be considered during the design. The most common way to get a picture of the sensitivity of each component is to do a Monte Carlo simulation. Every component is assigned its mean value, deviation and statistical distribution. Several simulations are done and with the limits for an accepted circuit performance given, it is possible to estimate the yield of the chip. This analysis, a so-called sensitivity analysis, can be done for pointing out the components which have the greatest impact on the response, but also to maximise the yield. This procedure is called design centring, which is a way of minimising the sensitivity for component variation. Since tuning is almost impossible for MMICs, and production costs are determined by the yield, it is important to get as many usable chips as possible out of each batch.

One misunderstanding concerning design of integrated circuits is that the specifications can be made previous to the design and that the designer's task is to construct a circuit that meets these specifications. A typical route when developing a new system or service is shown in figure 4.1. This figure does only show the small part of the route which involves design of ASICs. When the system configuration is considered, critical functions must be identified so that these functions can be realised as prototype circuits. This is necessary to make sure that the system under development is realisable. When these prototypes have been fabricated they must be thoroughly characterised to see if they meet the specifications. If not, the system must be reconfigured, which means that either must a new solution be developed or the specifications must be adjusted. This loop between system configuration and acceptance is a very important area for research. If this loop is lacking or badly performed, a bad solution may be the result. When the critical functions have been accepted, the system specifications can be fixed and the number of circuits and kind of technology are selected on a cost basis. The system can then be realised. The meaning of figure 4.1 is to point out that it is too late to involve ASIC designers when the system is to be realised.

5 Applications

The major advantage of MMICs is, as discussed in section 1.3, the low price per chip achievable at high production volumes. Thus, the applications for MMICs are foremost in markets where there is a demand for high volumes with a low price per chip. A market that exhibits these features, and thus is believed to be excellently suited for the MMIC approach, is the direct broadcast from satellite to domestic TVs, often called DBS (21). The frequencies involved are around 11-12 GHz, as well as 22 GHz. The circuits needed in this application, are front-ends for conventional parabolic antennas. These are low-noise amplifiers, oscillators, mixers, IF amplifiers and filters. Figure 5.1 shows a possible implementation of a DBS receiver. The noise performance of the input amplifier determines the size of the parabol which means that this is a critical element. Therefore, the front-end amplifier is sometimes made as a single amplifier low noise amplifier by utilising an ultra low noise transistor. This amplifier is made as a hybrid on a high quality substrate and tuned for minimum noise. The first stage can be considered as a part of the antenna and there is a trade-off between these two parts. Today, an ultra low noise p-HEMT transistor is usually used in the first stage. Such a transistor can have a noise figure in the order of 0.5 - 0.8 dB with an associated gain of 11 - 12 dB at 12 GHz. This reduces the noise requirements of the integrated LNA. Typical performance of a 3-stage LNA integrated in a 0.5 µm general purpose MESFET process is an NF of 3.5 - 4.0 dB and a gain of 15 - 20 dB at 12 GHz. The mixer can be implemented by using a dual-gate MESFET, which will give extra gain. Typical performance of such a mixer is an NF (single sideband) of 4 dB at 1 MHz off the carrier, a conversion gain of 6 - 10 dB and an IP_3 of 20 dBm. LO/RF isolation for such a mixer is considered to be good.

The local oscillator may be integrated on the chip as well. A very critical parameter for a local oscillator used in satellite reception is the phase noise. Generally speaking, MMICs have fairly low Q-factors. This results in parasitic oscillations and increased phase noise as compared to high-Q circuits. To make an MMIC useful for satellite reception, the phase noise has to be reduced. There are several schemes for achieving this. An important one for hybrid circuits, is the use of a dielectric puck in the feedback loop (22). Typical dimensions for such a dielectric cylinder with a resonant frequency of 12 GHz, is a radius and height of 2 mm (23). This is quite small when used in hybrid circuits, but normally too large when talking about MMICs.

Another way of reducing the noise, is by locking the oscillator's frequency to a very stable low-frequency crystal oscillator in terms of a phase-locked loop (24). A problem with this solution is the realisation of an MMIC frequency divider (24). This is needed in order to divide the frequency of the MMIC oscillator down to the frequency of the crystal oscillator. Such a frequency divider has, however, been developed (24), (25). This resulted in a 14 GHz oscillator exhibiting a noise level of -80 dBc/Hz at 1 kHz offset. This is considered to be sufficiently low for most communication systems (24).

When designing an MMIC for DBS it is advantageous to add an IFA to the chip to achieve extra gain, suppression of mixing

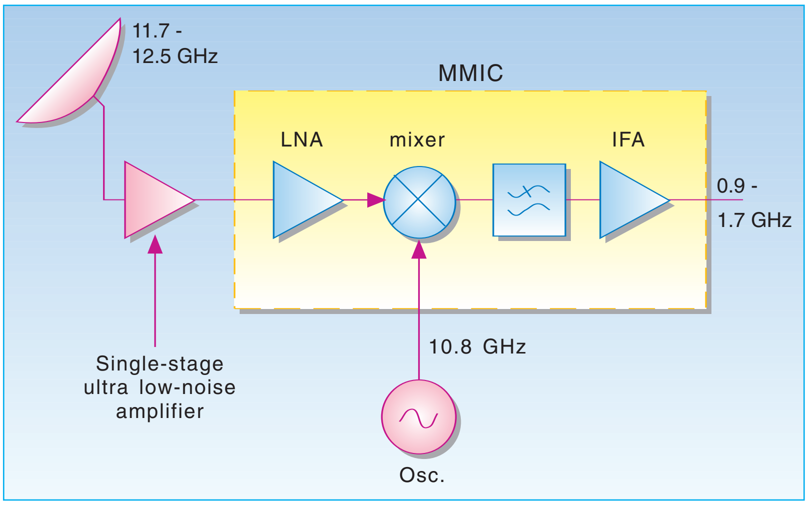


Figure 5.1 A possible implementation of a DBS receiver

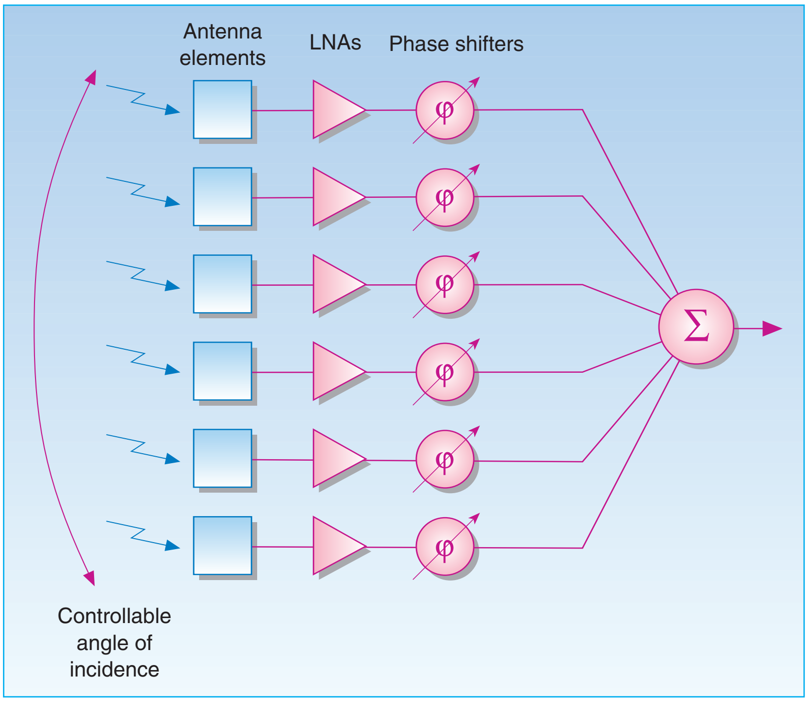


Figure 5.2 A phased-array antenna with readout electronics

products and increased output power. Typical performance of a 2 stage IFA is a gain of 10 dB at 0.9 - 1.7 GHz and an output power of 10 - 15 dBm. An IFA normalling has a bandpass response and works as a filter for mixing products. A complete chip containing these three functions is expected to be available from many manufacturers soon. Such a chip might cover an area of approximately 10 mm^2.

Another area of special interest for MMIC is the phased-array antenna. Such an antenna consists of several antenna elements. The signal received by each element is (often) amplified and then given a phase shift (see figure 5.2), and thus the antenna is able to receive signals from different directions depending on how these phase shifts are set. This application includes elements called phase-shifters (26), (27). Figure 5.2 shows a system including both the antenna arrays, which is often planar, and the readout electronics. In this application, area and power consumption becomes important as it involves a lot of chips. When using several (2) antenna elements it is obvious that hybrid circuits, which require soldering and tuning, is disadvantageous. The noise requirements of the LNA can be a reason for choosing a HEMT process for integration.

Another application is front-ends for optical communication (28) - (32). This includes both direct detection and coherent systems. To receive an optical signal, a photo detector is needed to convert the optical signal to a current. Sitting next to this is an amplifier, usually a transimpedance amplifier which converts the current into a voltage. One of the most profound features of amplifiers for optical receivers is the wide band. For instance, an amplifier for a coherent detection system has a band reaching from 50 MHz to 2.2 GHz (32).

It is envisaged that urban or suburban local area networks employing a point-to-multipoint configuration will become an important part of digital microwave communication systems (21). For this application, it is important that the outstations are designed with the potential to be manufactured at a low cost, and the use of MMIC technology should enable this large market to be realised. The frequencies involved would probably be above 20 GHz in order to achieve highly directional beams. 60 GHz would be particularly advantageous because the high atmospheric attenuation at this frequency may allow frequency reuse.

Cellular radio represents another low-cost/high-volume market. The frequencies are close to 1 GHz. The need for GaAs based devices in cellular radio units is significant, but confined to high efficiency transmitters, down-conversion MMIC subsystems and digital frequency synthesis modules (21).

Although it seems that hybrid circuits are advantageous for LNAs, the reality is not that simple. Consider, as an example, two parabolic antennas with a diameter of 55 cm and 90 cm, respectively. The 55 cm antenna has a gain of 35.5 dB with a noise temperature in the antenna and the interconnection of 31.9 K (33). The 90 cm antenna has a gain of 39.9 dB with a corresponding noise temperature of 29.2 K (33). These figures are at a frequency of 12 GHz. Now, consider three different realisations of the LNA based on state-of-the-art technologies (gate length 0.25 µm). A hybrid LNA with an ultra low noise transistor can achieve a noise figure as good as 0.8 dB. The corresponding noise figure for a monolithic LNA integrated in a HEMT process is typecally 1.6 dB (34) while it is approximately 2.6 dB for a monolithic LNA integrated in a MESFET process. The 55 cm antenna with the hybrid LNA will have a G/T = 15.94 dB/K. The 90 cm antenna with the p-HEMT LNA will have a G/T = 17.94 dB/K, and with the MESFET LNA G/T = 15.63 dB/K. A disadvantage of the smaller antenna is that it has a broader beam and thus may pick up interfering signals from adjacent satellites or terrestrial transmitters. For some satellites the small angle between adjacent satellites would determine the lower limit for the size of the antenna. This means that there is a trade-off between many different ratio, and a large antenna combined with a rather poor LNA can give the overall best performance/ratio factor. As seen from the above example a 55 cm antenna in combination with an LNA with NF = 0.8 dB has the same figure of merit as a 90 cm antenna in combination with an LNA with NF = 2.45 dB. The performance of the latter LNA is expected to be at hand for a straightforward standard MESFET technology in the near future. This means that this function can be integrated on the complete chip and comes almost for free.

6 Concluding remarks

The influence which the development of microelectronics has had on the development of new and improved telecommunication services is obvious and is a matter of no discussion. Both the complexity of these services and the reduced costs, which have increased the availability, is based on the possibility of integration of multifunctions on monolithic chips. We believe that the development of new and better microelectronic technologies will continue and further decrease the cost per function, improve the performance of these functions, and also make it possible to integrate completely new functions.

Microwave frequencies is a new area of applications for monolithic circuits. To date this area has been covered by expensive semiconductor devices and tubes. Monolithic integration has several advantages in the high frequency area such as shorter signal path, lower cost per function, small volume and area, and low power consumption. Price reduction opens up for a wider use while improved performance opens up for new applications. One interesting point is that when MMICs come into use in a high volume market with a potential for high profit this will be a major force behind the development of this technology. We believe that this is about to happen and that GaAs will be a very important technology through the 1990s. If the development as we have seen it the last 5 years continues in the same escalating rate, the very interesting 60 GHz frequency can be utilised in the near future. Technology with high f_t and high reproducibility is a condition for this to happen. GaAs is the semiconductor material with the best properties for use at microwave frequencies. Although silicon is moving upwards in frequency due to improvements in lithography, GaAs will benefit from the same degree of miniaturisation and stay ahead because of higher carrier mobility and electron velocity.

GaAs MMIC will become a very important technology by making it possible to utilise higher frequencies. It offers properties required for making functions with good performance at a low cost. Monolithic integration offers reduced costs by reducing the cost per function considerably. It is important to master this technology to be able to anticipate the development of new telecom services and new possibilities at higher frequencies.

7 References

1 Pucel, R A. Design considerations for monolithic microwave circuits. *IEEE transactions on microwave theory and techniques,* 29, 513-534, 1981.

2 Mehal, E, Wacker, R W. GaAs integrated microwave circuits. *IEEE transactions on microwave theory and techniques,* 16, 451-454, 1968.

3 Pengelly, R S, Turner, J A. Monolithic broadband GaAs FET amplifiers. *Electronics letters,* 12, 251-252, 1976.

4 Kato, H et al. A 30 GHz MMIC receiver for satellite transponders. *IEEE transactions on microwave theory and techniques,* 38, 896-902, 1990.

5 Sze, S M. *Semiconductor devices: physics and technology.* New York, John Wiley & sons, 1985. ISBN 0-471-83704-0.

6 Ladbrooke, P H. *MMIC design: GaAs FETs and HEMTs.* Boston, Artech House, 1988. ISBN 0-89006-314-1.

7 Tiwari, S. *Compound semiconductor device physics.* Boston, Academic Press, 1992. ISBN 0-12-691740-X.

8 Sivertsen, R, Paulsen, A. 283,000 cm^2/Vs mobility in GaAs grown by MBE using As_4. *Sixth European conference on molecular beam epitaxy and related growth methods, Tampere, Finland,* 1991.

9 Dingle, R et al. Electron mobilities in modulation doped semiconductor heterojunction superlattices. *Applied physics letters,* 33, 665-667, 1978.

10 Nguyen, L D, Larson, L E, Mishra, U K. Ultra-high-speed modulation-doped field-effect transistors: A tutorial review. *Proceedings of the IEEE,* 80, 494-518, 1992.

11 Foisy, M C et al. The role of inefficient charge modulation in limiting the current gain cutoff frequency of the MODFET. *IEEE transactions on electron devices,* 35, 871-878, 1988.

12 Lee, R E et al. Ultra-low noise millimeter-wave pseudomorphic HEMT's. *IEEE transactions on microwave theory and techniques,* 37, 2086-2092, 1989.

13 Duh, K H G et al. A super low-noise 0.1 μm *T*-gate InAlAs-InGaAs-InP HEMT. *IEEE microwave and guided wave letters,* 1, 114-116, 1991.

14 Nguyen, L D et al. 50-nm self-aligned-gate pseudomorphic AlInAs/GaInAs high electron mobility transistors. *IEEE transactions on electron devices,* 39, 2007-2014, 1992.

15 Goyal, R (Ed). *Monolithic microwave integrated circuits: Technology & design.* Boston, Artech House, 1989. ISBN 0-89006-309-5.

16 Feng, M, Laskar, J, Kruse, J. Super-low-noise performance of direct-ion-implanted 0.25-μm-gate GaAs MESFET's. *IEEE electron device letters,* 13, 241-243, 1992.

17 Feng, M et al. Ultra low-noise performance of 0.15-micron gate GaAs MESFET's made by direct ion implantation for low-cost MMIC's applications. *IEEE microwave and guided wave letters,* 2, 194-195, 1992.

18 Feng, M et al. Does the two-dimensional electron gas effect contribute to high-frequency and high-speed performance of field-effect transistors? *Applied physics letters,* 57, 1233-1235, 1990.

19 Kim, M E et al. GaAs heterojunction bipolar transistor device and IC technology for high-performance analog and microwave applications. *IEEE transactions on microwave theory and techniques,* 37, 1286-1303, 1989.

20 Wilson, K. Other circuit elements for MMICs. *GEC journal of research,* 4, 126-133, 1986.

21 Oxley, T H, Forrest, J R. The systems need for GaAs devices and integrated circuits. *GEC journal of research,* 4, 66-71, 1986.

22 Kajfez, D, Guillon, P (Eds). *Dielectric resonators.* Dedham, Artech House, 1986. ISBN 0-89006-201-3.

23 Heide, C F. *Monolithic microwave integrated circuits for communication systems.* Kjeller, Norwegian Telecom Research, 1990. (TF-report R38/90)

24 Ohira, T, Hiraoka, T, Kato, H. MMIC 14-GHz VCO and Miller frequency divider for low-noise local oscillators. *IEEE transactions on microwave theory and techniques,* 35, 657-662, 1987.

25 Ohira, T et al. 14 GHz band GaAs monolithic analogue frequency divider. *Electronics letters,* 21, 1057-1058, 1985.

26 Koul, S K, Bhat, B. *Microwave and millimeter wave phase shifters. Volume I: Dielectric and ferrite phase shifters.* Boston, Artech House, 1991. ISBN 0-89006-319-2.

27 Koul, S K, Bhat, B. *Microwave and millimeter wave phase shifters. Volume II: Semiconductor and delay line phase shifters.* Boston, Artech House, 1992. ISBN 0-89006-585-3.

28 Casao, J A et al. A comparison between three GaAs monolithic transimpedance amplifiers for optical communication systems. *Proceedings of the 22nd European microwave conference,* 330-335, 1992.

29 Petersen, A K, Jagd, A M. Tuned optical front-end MMIC amplifiers for a coherent optical receiver. *Proceedings of the 22nd European microwave conference,* 1065-1070, 1992.

30 Ishihara, N et al. A design technique for a high-gain, 10-GHz class-bandwidth GaAs MESFET amplifiers IC module. *IEEE journal of solid-state circuits,* 27, 554-561, 1992.

31 Darwazeh, I et al. GaAs MMIC optical receiver with embedded signal processing. *Proceedings of the IEE, part G,* 139, 241-243, 1992.

32 Olsen, T. 2.2 GHz dual-detector lightwave receiver employing current shunt feedback amplifier design. *Electronics letters,* 26, 510-511, 1990.

33 Morgan, W L, Gordon, G P. *Communications satellite handbook.* Wiley & sons, 1989.

34 Ayaki, N et al. Low noise MMIC amplifiers using HEMTs. *Electronics and communications in Japan, part 2,* 74, 1-8, 1991.

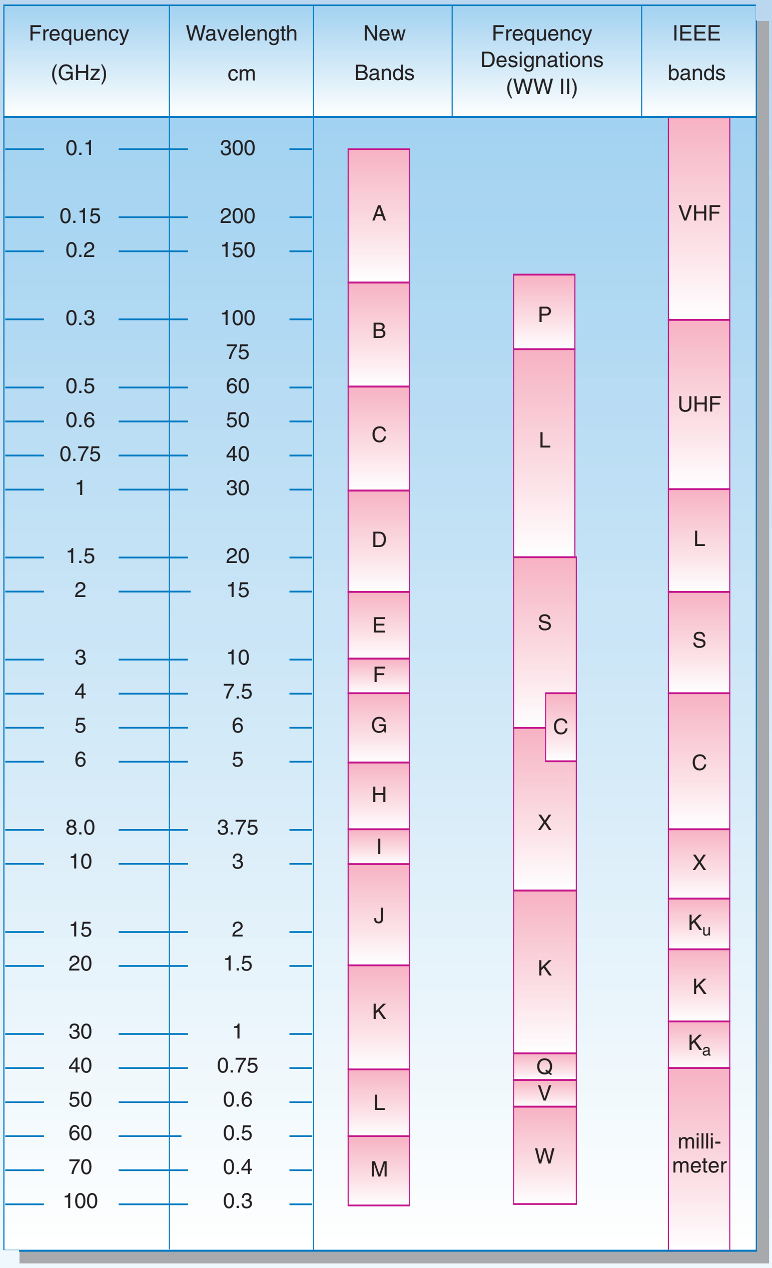
Acronyms and abbreviations

The following acronyms and abbreviations are widely used in conjunction with microwave components and circuits, both in this article and elsewhere.

ASIC	- Application Specific Integrated Circuit
FET	- Field Effect Transistor
HBT	- Heterojunction Bipolar Transistor
HEMT	- High Electron Mobility Transistor
HFET	- Heterostructure Field Effect Transistor
IF	- Intermediate Frequency
IFA	- Intermediate Frequency Amplifier
IP_3	- Third order Intercept Point
LNA	- Low Noise Amplifier
MBE	- Molecular Beam Epitaxy
MESFET	- MEtal Semiconductor Field Effect Transistor
MIC	- Microwave Integrated Circuit
MIM	- Metal Insulator Metal
MMIC	- Monolithic Microwave Integrated Circuit
MODFET	- MOdulation Doped Field Effect Transistor
MOCVD	- Metal-Organic Chemical Vapour Deposition
NF	- Noise Figure
p-HEMT	- Pseudomorphic High Electron Mobility Transistor
SDHT	- Selectively Doped Heterojunction Transistor
TDEG	- Two-Dimensional Electron Gas
2-DEG	- Two-Dimensional Electron Gas
TEGFET	- Two-dimensional Electron Gas Field Effect Transistor
VLSI	- Very Large Scale Integration

Microwave band designations

During World War II, the Allied Forces started to use letters to designate certain frequency bands. The letters were on purpose chosen without any logic so as to confuse the enemy. Unfortunately for novices, these letters are still used for frequency band designations. To make the confusion complete, there are in use other band designations as well. As a consequence, the same letter can designate somewhat, or completely, different frequency bands in the different "standards". In the figure below, we have shown the band designations from World War II, together with those of IEEE and a new system defined by the US Department of Defence. The IEEE designations are the ones in most widely use.



File transfer performance tuning in X.25 and TCP/IP

BY PER SIGMOND

681.327.8:006

1 Introduction

Public packet switched data network services based on the X.25 network protocol [1] exist in most countries around the world today. X.25 networks give reliable and error-free connections, multiple sessions, and distance independence. For many customers it is the only affordable option available for long distance file transfers or LAN interconnection. Subscribers can choose from different signalling speeds ranging from about 2400 bits/second to 64000 bits/second or more. Compared to the services offered by competing solutions (such as modems over telephone-lines) X.25 is often the better choice. The protocol has even found its way into new digital public services like the ISDN.

In spite of these good qualities X.25 networks are often said to be "slow". The subscribers feel that they do not get the performance they expect compared to the physical signalling speed they pay for. And very often they are right! Is this because X.25 is a lousy network protocol or because the PTTs do bad network dimensioning? Probably not. I think the reason very often is badly tuned X.25 protocol parameters. My feeling is that both subscribers and PTT officials know too little about the impact of parameters like *window size* and *packet size*. As this article shows, changing these parameters can in some cases increase throughput more than 5 times! But it can also bring some unwanted side effects in protocols run on top of X.25 (we have investigated TCP/IP [2] [3]). This article show how to control these effects and what benefits you can get from tuning the X.25 parameters. The results presented are based on file-transfer measurements done in the Norwegian Datapak network by the project "Performance measurements over X.25" [4]. This project was performed by Agder Research Foundation and ITS, and was funded by Norwegian Telecom Research.

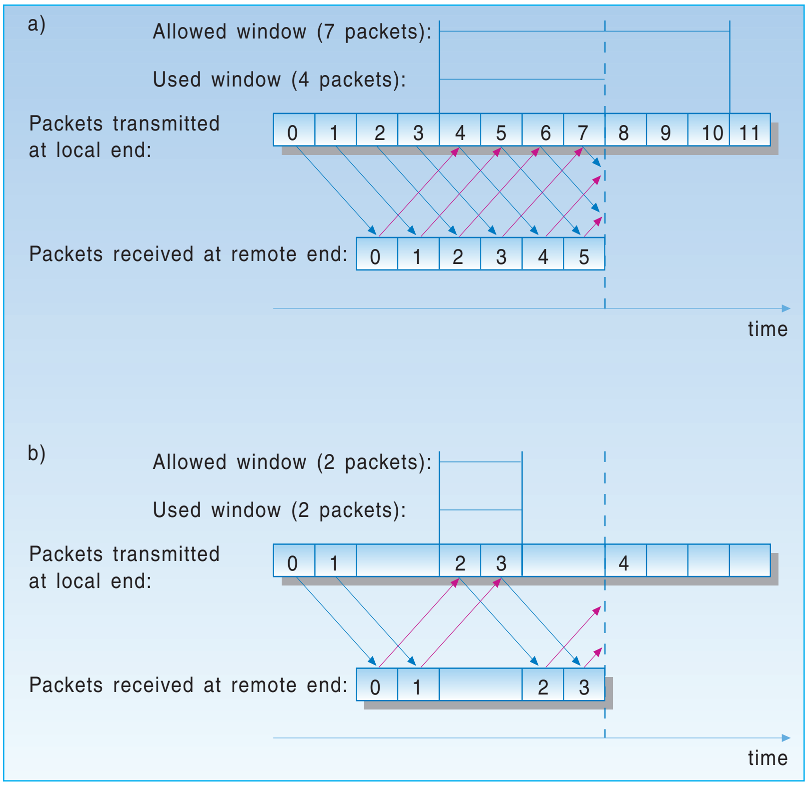


Figure 1 Flow control using "sliding window" with window sizes of a) 7 packets and b) 2 packets

2 Protocol parameters

2.1 Window size

Computers exchange information by transmitting packets of data according to specific rules known as protocols. Connection oriented protocols such as X.25 and TCP must (among other things) have a way to control the flow of these packets (when to transmit them). This is done by a mechanism known as *sliding window* together with *positive acknowledgement*. Basically it works like this:

> When the transmitter starts it is allowed to transmit an amount of data immediately without getting any acknowledgements from the remote receiver. This amount of data is known as the window size. The transmitter keeps track of how much of the window it is currently using; that is the amount of data it has sent that is not acknowledged. If this "used window" reaches the (allowed) window size it must stop transmitting until it receives acknowledgements on the data first sent. Upon receiving such acknowledgement the transmitter moves the window onward and is free to send some more data. In this way the window "slides" across the file.

Figure 1 a) shows a situation where the transmitter never reaches the window size limit, while in b) the window size is so small it has to stop waiting for acknowledgements, resulting in decreased efficiency (only 3 packets sent so far).

The X.25 protocol counts the window size in *packets* while TCP counts it in *bytes*. For X.25 layer-3, the Datapak allows window sizes from 2 packets to 7 packets (2 is the default). TCP window sizes can vary from a few bytes to several thousand bytes. Common values in UNIX systems are 4096 bytes and 8192 bytes.

2.2 Packet size

The packet size is defined as the amount of user data you are allowed to stuff into one protocol packet. The packet sizes allowed in Datapak are 128, 256, 512 and 1024 bytes. The default is 128 bytes. TCP is less accurate at this point, but common values are 512 bytes and 1460 bytes (fits into one Ethernet frame).

3 X.25 performance measurements

3.1 File transfer performance

To show the impact of X.25 protocol parameters on file transfer performance, we will show some measurements done in the Norwegian public Datapak network. The set-up includes transmitting big files over long distance going through several X.25 routers in the Datapak. The signalling speed at both ends were 64000 bits/s. File transfers were done for all combinations of X.25 packet size and window size. Each value is based on 5 file transfers of 500 kilobytes each.

Figure 2 shows file transfer performance dependant on the protocol parameters window size and packet size for our test-connection. As mentioned before the default values of Datapak subscriptions are window size 2 and packet size 128, and these are the settings most people use. We can see that performance is increased by 250 percent just by increasing the window size to 7. Performance is again doubled if we increase the packet size to 1024 bytes. This means a *performance improvement of more than 500 per cent* merely by adjusting protocol parameters at both ends! In Datapak you pay a small fee for the ability to negotiate these parameters, but for file transfer it seems to be worth the price!

It is hardly surprising that bigger packets give better performance. After all you get more data into each packet and therefore less protocol overhead in terms of headers and acknowledge packets. But why is the performance so heavily dependant on the window size? To understand this we have to look at another aspect of the network; the transmission delay.

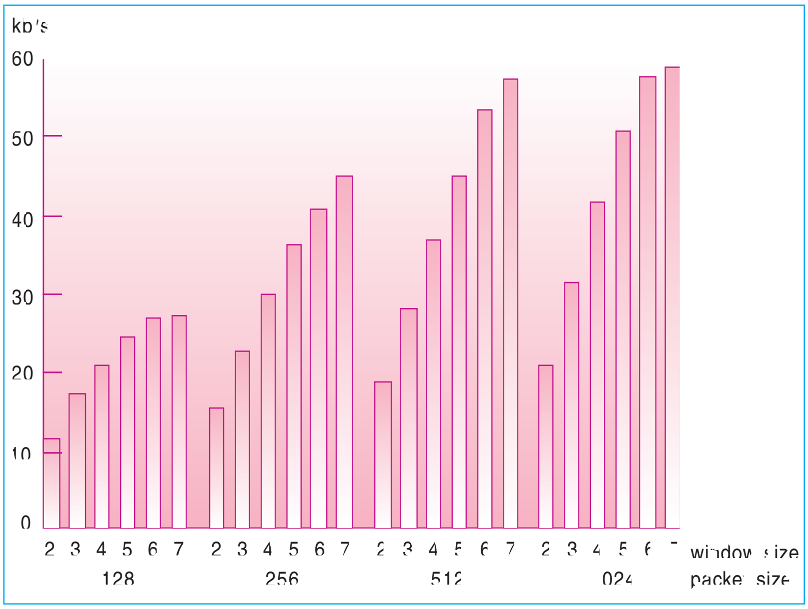


Figure 2 File transfer performance using X.25

3.2 Transmission delay

Transmission delay is defined here as the time for one X.25 packet to travel from the transmitter to the receiver. Figure 3 shows the delays for our test connection. As you can see, the delay is heavily dependant on the size of the packet. The main reason for this is the way the intermediate nodes (routers) in the X.25 network exchange packets; the connection oriented layer-2 protocol. For this protocol to perform its task, each packet has to be clocked in to the node in its entirety and acknowledged upon before it can be transmitted further. The clocking of course takes more time the longer the packet is. This way each router in the network function as a store-and-forward node. With many intermediate nodes in the communication path this could mean long delays.

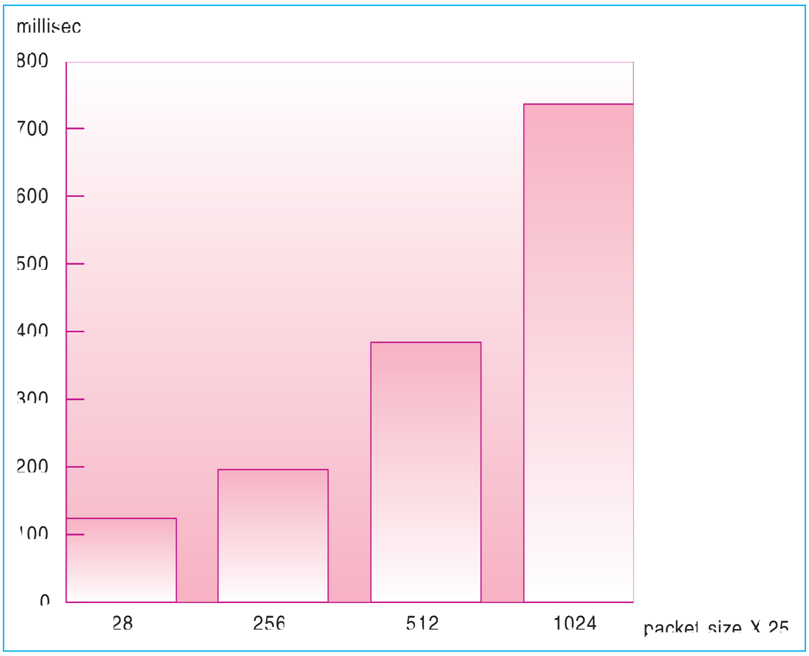


Figure 3 Transmission delay

What impact does this delay have on the protocol performance? In fact it is the reason why sliding window is used in the acknowledgement algorithm. The time between packet sent and acknowledge received is proportional to the transmission delay. Set up with a small window size the transmitter very likely will reach the window limit before it receives any acknowledgements. This means that it has to stop transmitting and thereby waist bandwidth (Figure 1). For optimal performance the window size must be large enough for the transmitter never to reach the window limit, which means it never has to stop transmitting waiting for acknowledgements.

4 TCP/IP over X.25

4.1 TCP performance measurements

When interconnecting LANs over X.25 connections we also have to consider how higher level protocols influence the performance. There are several ways to encapsulate LAN datagrams into X.25 packets, and for TCP/IP the most frequent one is the RFC877 [5]. This specification says how IP routers should use X.25 connections for transferring IP datagrams. When the IP packet is too big to fit into one X.25 packet it is just chopped up into smaller parts and reassembled at the receiving side of the X.25 connection.

From what we learned from the X.25 case we could expect the TCP/IP measurements to follow the same pattern with respect to X.25 protocol parameters. But as the TCP/IP figures show we also have to consider the TCP window size. In fact the transmission delay introduced by big X.25 packets really breaks TCP performance when using default TCP window values. To show this we have included two measurement series with different values of the TCP window size; 4096 bytes (the default on our machines) and 25000 bytes. The results are based on 5 file transfers of 500 kilobytes each.

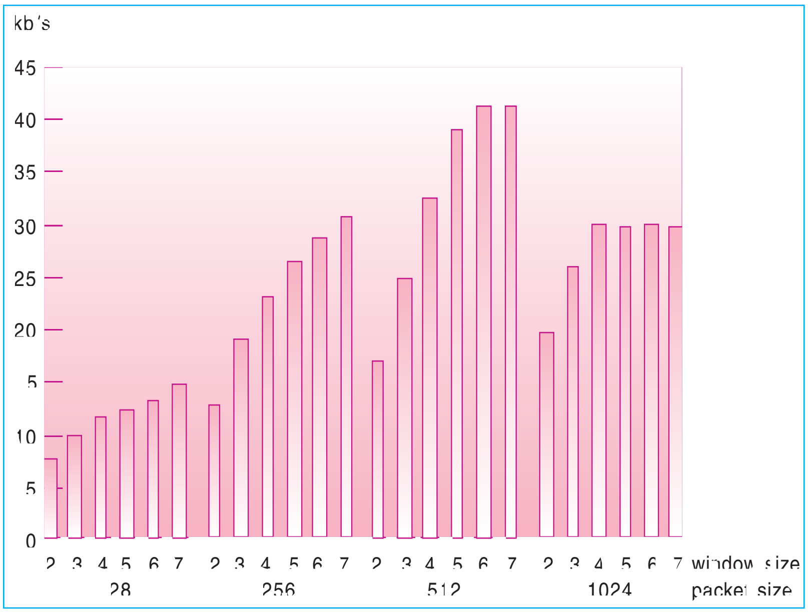


Figure 4 File transfer performance using TCP/IP over X.25 with TCP window size 4096 bytes

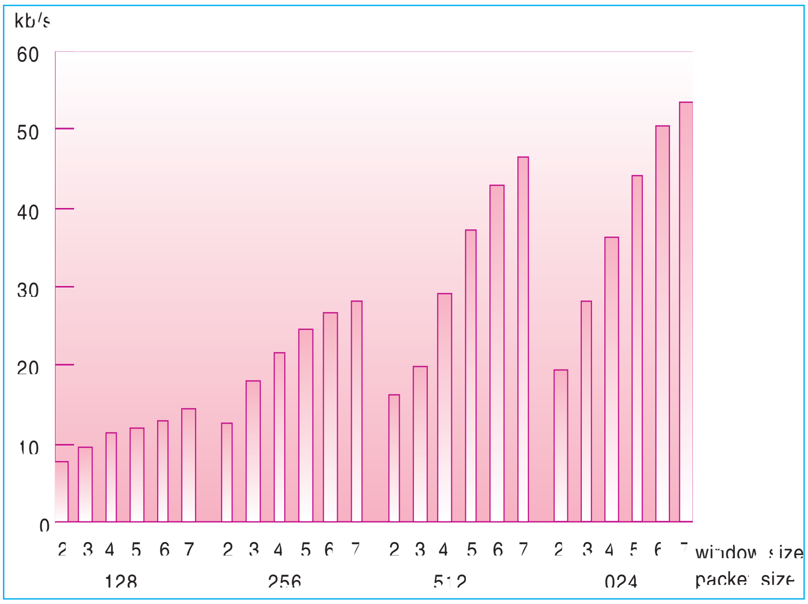


Figure 5 File transfer performance using TCP/IP over X.25 with TCP window size 25000 bytes

4.2 The Delay Bandwidth Product

Looking at Figure 4; why is TCP performance so poor for X.25 packet size 1024? To understand this we have to introduce a new number; the *Delay Bandwidth Product (DBWP)*. For our purpose we define this number as the product of the round trip delay of a TCP packet and the offered bandwidth. The *DBWP* then becomes an estimate of how much data the TCP transmitter can send into the transmission channel before the first acknowledge arrives from the remote end. Given this number we can calculate the minimum size of the TCP window.

To measure the *DBWP* we first do the assumption that the round trip delay of a TCP packet is the transmission delay of a long packet plus the transmission delay of a short packet (true for file transfer). To measure this we can use the UNIX utility "ping" which sends and receives ICMP [6] echo-packets and measure their round trip delay. The TCP round trip delay TCP_{delay} can then roughly be expressed as

$$TCP_{delay} = (D_{long} / 2) + (D_{short} / 2) \qquad (1)$$

where D_{long} is the measured round trip delay in seconds of a "long" (e.g. 1024 byte) ICMP packet and D_{short} is the same for a "short" (e.g. 128 byte) ICMP packet.

Secondly we assume that the offered bandwidth $X.25_{rate}$ is equal to the measured X.25 transfer rate for the given connection. If we express the $X.25_{rate}$ in bits/second and the DBWP in bits, we get the following equation:

$$DBWP = TCP_{delay} * X.25_{rate} \qquad (2)$$

4.3 The TCP window size

We now have a base for estimating the necessary TCP window size TCP_{win}. But first we have to consider how most TCP implementations make use of the window. In order to avoid a condition called "the silly window syndrome", TCP implementations actually use only about 75 % of the offered window [7]. Furthermore we assume that we get optimal performance when the TCP transmitter never reaches its window limit; that is 75 % of TCP_{win}. If we express TCP_{win} in bytes and DBWP in bits, we get that optimal performance is achieved when:

$$DBWP \leq TCP_{win} * 8 \text{ bits/byte} * 0.75 \qquad (3)$$

$$TCP_{win}\ [bytes] \geq DBWP\ [bits] / 6 \qquad (4)$$

This equation (4) is based on many assumptions, and should only be taken as a rule of thumb, but given all the different implementations of the TCP protocol it is difficult to be more precise.

To investigate how these assumptions would hold for our set-up we made some control measurement series keeping the X.25 conditions constant while adjusting TCP_{win}. The two curves in Figure 6 show the performance for X.25 packet sizes 512 and 1024 bytes. X.25 window size was held at 7 packets for both. We measured the *DBWP* to be 35600 bits and 53400 bits giving a TCP_{win} of 5933 bytes and 8900 bytes. As Figure 6 shows, our "rule of thumb" (4) gives a good estimate of the optimal TCP_{win}.

5 Practical conclusions

5.1 X.25 connections

We have seen that tuning protocol parameters for the X.25 connection can give tremendous improvements on file transfer performance. This is especially true if you communicate over long distances and/or use high signalling speeds.

Are there any drawbacks from adjusting the parameters? Very few as far as I can see provided your network equipment has enough buffer space to handle it.

5.2 TCP/IP over X.25

When using TCP/IP over X.25 there is also a lot to gain by adjusting the X.25 protocol parameters. But one must be aware of the increased *DBWP* (2) one can get by increasing the X.25 packet size. The remedy is of course to increase the TCP window size accordingly, but this is not always possible. In many cases it involves patching the operating system of your machine or changing the source code of the communication programs. The practical solution will therefore often be to leave the TCP window as it is. The best choice is then to set the X.25 window size to the maximum value, and adjust the X.25 packet size upwards until you reach the *DBWP* your TCP window can handle.

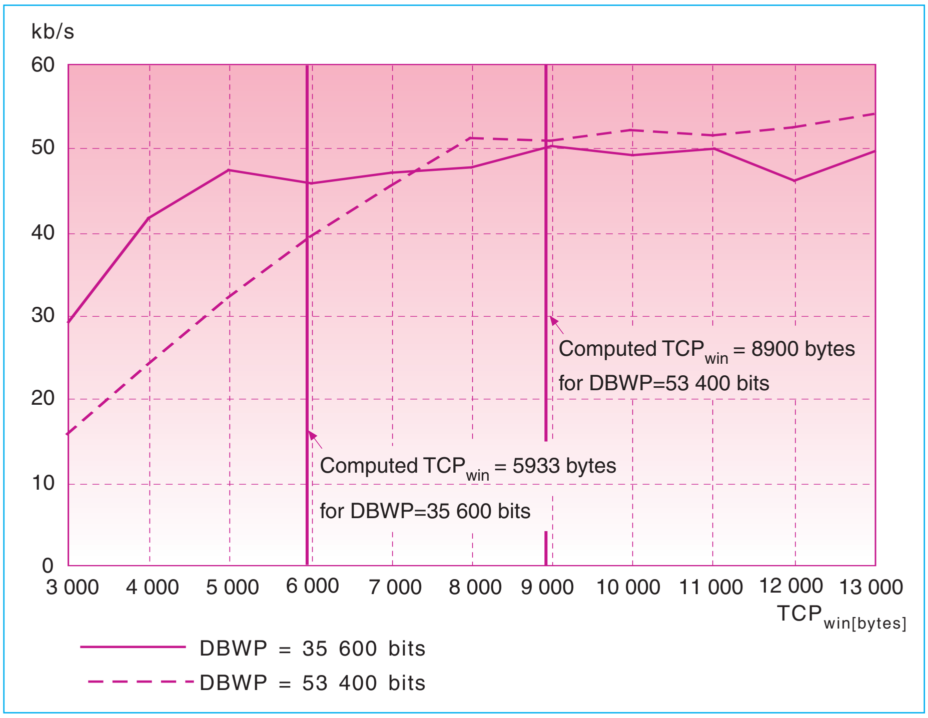


Figure 6 File transfer performance vs. TCP window size TCP_{win} for two different values of DBWP

6 References

1 CCITT X.25. *Interface between data terminal equipment and data circuit terminating equipment for terminals operating in the packet mode on public data networks,* CCITT Recommendation X.25, 1980, 1984, 1988.

2 RFC 791. Postel, J, (ed.). *Internet protocol - DARPA Internet program protocol specification,* RFC 791, September 1981.

3 RFC 793. Postel, J, (ed.). *Transmission control protocol - DARPA Internet program protocol specification,* RFC 793, September 1981.

4 Sigmond, P, Elvheim, Sorteberg, I. *Performance measurements over X.25.* Kjeller, Norwegian Telecom Research, 1992. (NTR paper N31/92.) Language: Norwegian.

5 Korb, J T. *A standard for the transmission of IP datagrams over public data networks.* RFC 877, September 1983.

6 Postel, J. *Internet control message protocol - DARPA internet program protocol specifications.* RFC 792, USG/Information Sciences Institute, September 1981.

7 Clark D D. *Window and acknowledgement strategy in TCP.* RFC 813, July 1982.

Software maintainability – on purpose or accidentally?

BY MAGNE JØRGENSEN

681.3.06.004.5

Abstract

In spite of an enormous productivity increase among software maintainers the maintenance problem is increasing. One reason for this may be the fact that it is common to believe that "good programmer practice" and "modern development methods" alone will give maintainable software. This is a dangerous belief and this article argues for an extensive and problem specific maintenance preparation. An example and an outline for doing this are presented.

"The first step (to face the phenomenon of software system change), is to accept the fact of change as a way of life, rather than an untoward and annoying exception." [1)]

1 Introduction

During the early days of programming a normal maintenance rate was four boxes of cards per full time maintenance programmer, (16). A box of cards contained about 2,000 cards and each card one line of assembly code. It takes about three assembly language lines of code to produce the functionality of one Cobol line of code (6). A full time maintenance programmer was consequently able to maintain the functionality of $(4*2000)/3 \cong 2700$ lines of Cobol code.

Times have changed! In a study I conducted at the system development section (ES) at the Norwegian Telecom in 1991/92, see (7), I found the average number of lines of Cobol code maintained per full time maintenance programmer to be more than 120000. That is about 45 times more. With the same productivity as in the "good old days" and the same number of lines of code as today we probably would have no unemployed people in Norway. The system development section referred to above alone would require more than 1000 full time maintenance programmers to maintain its 2.7 million lines of code (see figure 1).

In spite of this enormous productivity increase I believe software maintenance should be considered a most serious problem in the software area today. The facts and opinions presented in *box one* support this belief.

This article outlines one approach to decrease the maintenance problems: preparing the software for software maintenance. Using this approach the maintainability characteristics of the software will be based on rational decisions instead of trusting that "good software development practice" will result in desired maintainability.

2 What is software maintenance?

A search in the literature for definitions of software maintenance may be a bit confusing, since there exist many quite different and even slightly contradictory definitions and descriptions of the term. In addition there are several other terms denoting almost the same. Even within a homogenous programming team maintaining the same software, rather different interpretations may be present, see (7). Examples of different software maintenance definitions and synonyms are collected in *box two*. In many contexts it is very important to be aware of this "confusion" to avoid misunderstandings.

I do not think software maintenance is a good term, since maintenance normally is associated with physical degradation and routine work. Nevertheless, I will use the term 'software maintenance' here, since the introduction of a new term, like 'further development' may introduce misinterpretations as well.

There is no "best" definition of software maintenance. Personally, I prefer definition 2, box two, when speaking generally about software maintenance, as I will do in this article. That

[1)] Frederick P Brooks Jr., *The mythical man-month*, p 117.

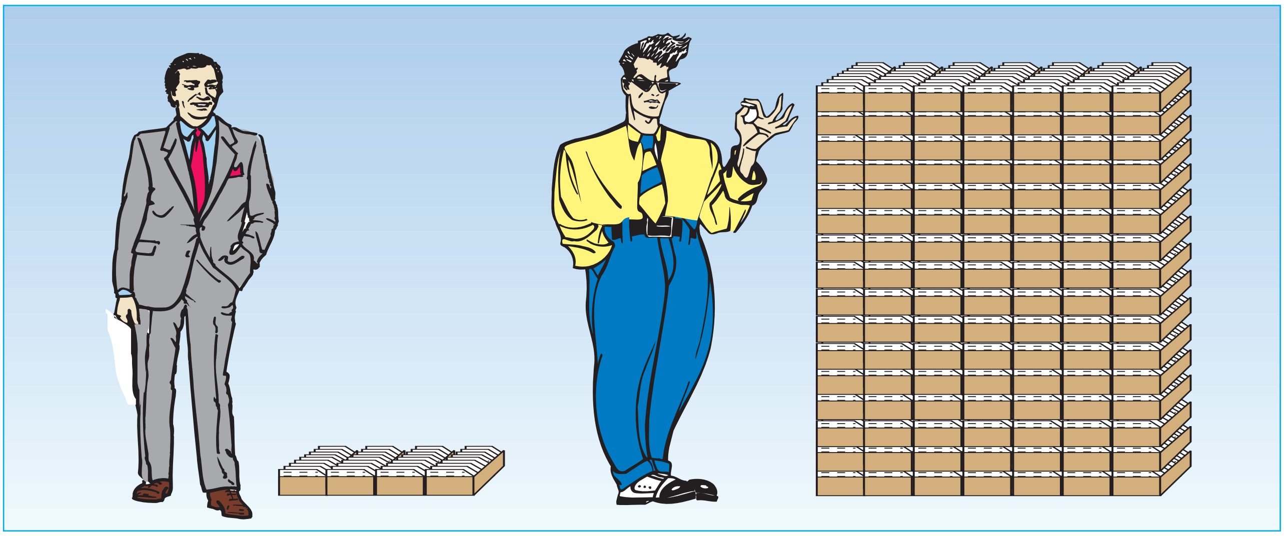


Figure 1

definition is not very "operational". Thus, when conducting an experiment another definition will probably be more useful, for instance definition 1 or 6, box two.

3 Preparing for maintenance

Preparing for maintenance should take place in every phase of the software life cycle. The objective of the activity is to make the subsequent maintenance more efficient. The activity is different from for instance software testing, which aims at removing faults in the software, not at more maintainable software.

As indicated in box one, the maintenance costs constitute a large part of the total software costs. It would, for this reason, be natural that activities aiming at more maintainable software would be important in most software development projects. I do not think it is, today. I believe it is rather unusual to make an extensive analysis of how to develop software to achieve the desired maintainability. This may be due to time pressure or lack of knowledge about what makes software maintainable, but also due to the two beliefs described below:

- *Software easy to develop will be easy to maintain.* That is, take care of the development and the maintainability will take care of itself.
- *Choice and use of "modern" software development methods are sufficient preparation for the maintenance of software.*

Of course, none of these beliefs are totally unsupported by empirical evidence. The productivity increase among software maintainers, described in the introduction, may, to a large extent, have been caused by "modern" development methods, like the method of structured programming. However, the beliefs are not always correct today and may be even less correct in the future. This is further illustrated in the next two sections.

3.1 The impact of the development method on the maintainability

When someone, for instance a vendor of software development methods, claims that his method produces more maintainable software than other methods, two assumptions seem underlying:

- *Assumption one:* The choice of development method has a considerable influence on the maintainability of the software. (The method would otherwise be of little interest for the maintenance programmers.)
- *Assumption two:* All "intended" ways of using the method produce more maintainable software than other methods faced with the same problem to solve.

Most programmers probably would accept assumption one, although we often do not know! It may be dangerous to accept apparently obvious beliefs without having enough supporting facts, or to quote Sherlock Holmes: *"It is a capital mistake to theorise before one has data. Insensibly one begins to twist facts to suit theories, instead of theories to suit facts."*[2)]

Characteristics not part of a method, for instance the experience and "intelligence" of the programmers, may be the far most dominant determinators of the maintainability.

Newel and Simon, in (12), propose that *"The task environment (plus the intelligence of the problem solver) determines to a large extent the behaviour of the problem solver ..."* This proposition is based on a number of experiments where moderately difficult problems in chess, symbolic logic and cryptarithmetic puzzles were solved, and the proposition might support assumption one. We should, however, have in mind that software development and maintenance probably are more "compound" activities than the activities studied by Newel and Simon.

Surprisingly, there has been very little empirical research on the impact of development methods on the maintainability of software. In addition, the few research results we have are not unambiguous, see for instance (3). This lack of research has resulted in the unsatisfactory situation where the choice of development method cannot be based on "empirical facts". The choice of development method, programming languages and tools is often more like an "act of faith" than a "scientific" decision.

Assumption two is obviously dubious. It is, for instance, easy to construct examples where an unstructured C-program with GOTOs looks, in my view, more maintainable than a well structured[3)] C-program solving the same problem, see *box three*. This means that even the assertion that structured programming produces more maintainable programs than unstructured programming with GOTOs is not always correct!

The almost total avoidance of GOTO-constructs in many programs developed in the last few years is hence not only a good sign. There may be a situation where the purpose, that is increase software maintainability, of a software engineering rule is forgotten and the rule itself is given the status of a principle; "You shall not use GOTOs". The situation for structured programming, object oriented programming and for development methods is much the same.

3.2 The profitability of preparing for maintenance

Software maintenance may be divided into maintenance dealing with error corrections and maintenance dealing with other program modifications, e.g. enhancements. The first category, corrections of errors, only accounts for around 20 % of the total maintenance costs, according to (8) and (13). 80 % of the software maintenance cost is consequently caused by modifications not related to error corrections.

It is commonly accepted that an error detected in the design or requirement phase normally costs several times less to correct than the same error detected in later phases, like the system test

---

2) Sir Arthur Conan Doyle, *A scandal in Bohemia.*

3) There is a lot of confusion about structuredness. Here I use the definition from(8) where a structured program is defined as a program with "Essential complexity" = 1.

phase or the maintenance phase. The German company Siemens has calculated that the cost of correcting an error in the operation system BS2000 in average amounts to about 2,000 DM in the development phase, 6,000 DM in the system test phase and 20,000 DM in the maintenance phase, that is about ten times more in the maintenance phase than in the development phase.

Do we have a similar relation between modifications prepared for the development phase and modification not prepared for? If we have, it will be profitable to be very concerned with preparations for modifiability. No study, as far as I know, has focused on this question.

It is, however, rather obvious that a vendor must release new program versions quickly, with high quality and low costs to stay in business and to generate profit. This requires a highly maintainable software, I think.

There are arguments against the profitability of planning for software modifications, as well. A major argument is that if it really was profitable to plan for future modifications the software companies probably would have been doing it. The software companies invest a lot of money in the software and want the best possible return on their investment. For this reason the software will, assuming rational investors, adjust to an optimum maintainability, given the time, personnel, and the money available.

Another argument against the planning is that we do not know what the future modifications will be. Preparing for future modifications will thus be planning for the unexpected, a meaningless activity.

I think the arguments against planning for modifiability are not valid and in section 3.3 I give argument supporting my opinion. I do this in an outline of how to do the maintenance preparation.

3.3 How to do maintenance preparation

As said earlier, the modification part of software maintenance consumes about 80 % of the total software maintenance costs. For this reason, I will mainly focus on the preparation for modifications in this chapter. I will first give an example of maintenance preparation, section 3.3.1, and then give a brief outline of a possible maintenance preparation approach in section 3.3.2.

3.3.1 Example of maintenance preparation

Let us say we want to develop a tax calculation program, where the tax rates are described as follows:

Tax rate	*Income limits*
0 % of the first	122,000
10 % of the next	36,000
17 % of the remaining net income	

The calculation of the tax can be implemented in several quite different ways, see *box fou*r for six different implementations in a Pascal-like language. A programmer not preparing the implementation for subsequent maintenance will probably implement the tax calculation the way he intuitively feels is the best and in accordance with "good programmer practice". For me this was alternative 1 which, after some analysis, turned out to be one of the least maintainable alternatives.

Typical changes in tax rules are changes in the tax rates, the income limits for the different tax rates and the number of tax rates.

Change of tax rates and income limits can be easily performed on all alternatives, except change of the first tax rate (0 %)! Alternatives 1, 2, 3, and 6 have not prepared for a change of this tax rate. If it is likely that this rate will remain equal to 0 %, it is OK. If not, we should choose another way to implement the tax calculation.

Changes of the number of tax rates, for instance by introducing a new tax rate and income limit as indicated below, involve significantly more difficult changes for alternatives 1, 2, and 3 than alternatives 4, 5, and 6.

Tax rate	*Income limits*
0 % of the first	122,000
10 % of the next	36,000
17 % of the next	50,000
20 % of the remaining net income	

Box five illustrates the difference in the modifiability between alternatives 1 and 5. The modification of alternative 5 is much more difficult and error prone than the same modification of alternative 1.

Therefore, if it is probable with an introduction of one or more tax rates, alternatives 4, 5, and 6 should not be chosen.

3.3.2 Outline of an approach

The approach may be divided into two parts:

- Part one: Find the most probable and critical changes of the software.
- Part two: Make choices enabling an easy modification of the software for those changes.

The first part consists in "looking to the future" based on studies of what the typical changes have been, studies of what typical changes of similar software have been and "well-qualified guesses". The objective is to find the most probable and critical changes of the software, either forced by functionality change, property change (performance, use of memory, etc.), or environmental change of the software.

Questions typically to be answered:

- Will the hardware, programming language and tools probably have to change during the life of the software?
- To which degree is the capacity, e.g. memory, hard-disk, performance, likely to increase/decrease?
- To which degree will the user interface probably have to be changed?

- What is the most probable and critical extensions to the software?

The result from part one makes it possible to divide the general maintainability objective, that is "the software should be maintainable", into more manageable maintainability objectives, like "the change of tax rate in the software should be simple". Whenever possible, the objectives should be controllable, like "all hardware dependent instruction should be isolated and commented" (in order to know what to change when the hardware is getting changed).

There is no *general* approach for enabling an easy modification as desired in part two. I believe, however, that very often there are approaches for enabling easy modifications in *concrete* cases, like the one indicated in the previous example, see section 3.3.1.

The approach should be applied in each of the software life cycle phases.

References

1 Bennet, K H. The software maintenance of large software systems: management, methods and tools. *Reliability engineering and system safety,* 32, 135-154, 1991.

2 Boehm, B W, Papaccio, P N. Understanding and controlling software costs. *IEEE Transaction on software engineering,* 11, 1462-1477, 1988.

3 Boehm-Davies, D A et al. The role of program structure in software maintenance. *International journal of man-machine studies,* 1, 21-63. 1982.

4 Freedman, D P, Winberg, G M. A check-list for potential side effects of a maintenance change. *Techniques of program and system maintenance,* 61-68, 1980.

5 IEEE/ANSI. *An American national standard IEEE standard glossary of software engineering terminology.* ANSI/IEEE Standard 729, 1983.

6 Jones, C. *Applied software measurement.* New York, McGraw-Hill, 1991. ISBN 0-07-032813-7.

7 Jørgensen, M. *Vedlikeholdbarhet til programvare: Forstudie ved seksjon for systemutvikling (ES) i Televerket.* Kjeller, Norwegian Telecom Research, 1992. (TF paper N29/92.)

8 Lientz, B P, Swanson, E B. *Software maintenance management.* Reading, Massachusetts, 1980. ISBN 0-201-04205-3.

9 McCabe, T. A complexity measure. *IEEE Transaction on software engineering,* 3, 312-327, 1976.

10 Munson, J B. Software maintainability: A practical concern for live-cycle costs. *IEEE Computer,* 103-109, November 1981.

11 Foisy, M C et al. The role of inefficient charge modulation in limiting the current gain cutoff frequency of the MODFET. *IEEE transactions on electron devices,* 35, 871-878, 1988.

12 Newell, A, Simon, H A. *Human problem solving.* New Jersey, Prentice-Hall, 1972. ISBN 13-445403-0.

13 Nosek, J T, Palvia, P. Software maintenance management: Changes in the last decade. *Journal of software maintenance,* 2, 157-174, 1990.

14 Ogdin, J L. Designing reliable software. *Datamation,* 18, 71-78, 1972.

15 Riggs. Computer system maintenance. *Datamation,* 15, 227-235, 1969.

16 Vick, C R (ed). *Handbook in software engineering.* New York, Van Nostrand, 1984. ISBN 0-442-26251-5.

Box one

Facts and opinions about software maintenance

- More than 50 % of all software expenditure goes to software maintenance. This rate has increased in USA from 49 % in 1979 (9) to 58 % in 1990 (14).
- 2 % of the GNP in USA is spent on software maintenance (1).
- Maintenance problems are much the same as during the 1970's, despite advances made in structured methodologies and techniques (14).
- The size of the program is increasing. In 1970 the average size in USA was 48000 lines of code (9) and in 1990 204000 (14). (8) measured an average length of 120000 lines of code per program system at the Norwegian Telecom.
- The age of the programs is increasing. In 1979 the average age in USA was 4.75 (9) and in 1990 5.5 years (14). NB: The surveys did only count programs older than one year and the real average will hence be somewhat lower. (8) measured an average of 4.5 years on software at the Norwegian Telecom.
- Software quality standards, like ISO 9001/9000-3 do only prescribe the way software should be developed to get the desired quality. They do not describe how the quality, for instance the maintainability, of the product should be verified.
- In 1980 75 - 80 % of existing software was produced prior to significant use of structured programming, not to mention object oriented programming (4). Much of that software is still in operational use and maintained.
- The software costs are rapidly increasing, about 12 % per year (2). This means that the software maintenance costs are getting doubled every 6th year; assuming a constant rate between total software costs and the software maintenance costs.
- "The demand for new software is increasing faster than our ability to develop it. The reason for this is often a limited supply of personnel and funding, much which must currently be devoted to supporting the evolution of existing software" (that is, software maintenance) (2).
- "The programmers making the maintenance are usually less experienced than the implementors" (11).
- "... modifying software is a complex, error-prone process; maintenance documentation is inadequate; testbed resources are limited; and the people involved in maintenance are a different crew from the developers and generally have little understanding of the software structure" (11).

Box two

Software maintenance definitions

1 (5): The modification of a software product after delivery to correct faults, to improve performance or other attributes, or adapt the product to a changed environment.

2 (1): The activities (technical and managerial) that are undertaken on the software subsequently to the development. The development is complete when the product is delivered to the customer or client, and the software installed and released for operational use.

3 (14): The continuing process of keeping the program running, or improving its characteristics.

4 (15): The activity associated with keeping operational computer systems continuously in tune with the requirements of users, data processing operations, associated clerical functions, and external demands from governmental and other agencies.

5 (8): All modifications made to an existing application system, including enhancements and extensions.

6 (7): Modifications and error corrections made to an existing application system, including introduction of new functionality demanding less than 1 - 2 man-months programming.

7 Bug-fixing (colloquial).

Software maintenance synonyms

Enhancement, Support, Further development, Production programming, Phase 2, Software evolution, Post deployment software support (PDSS), Software renewal, System redevelopment, Software update and repair.

Box three

The two C-programs below insert ten tuples (t[0]...t[9]) into a database.

If an insert fails a rollback is performed and the program is stopped.

Program 1:

Structured version without GOTOs.

```
main()
{
    ......
    i = 0;
    status = OK;
    while((i < 10)and(status == OK));
    {
        status = insert_tuple(t[i]);
        i++;
    }
    if(status <> OK)
    {
        rollback;
        exit(FAIL);
    }
    else
    {
        ......
    }
}
```

Program 2:

Unstructured with GOTOs.

```
main()
{
    ......
    for(i = 0; i < 10; i++)
      if(insert_tuple(t[i]) <> OK)
        goto rollback_and_exit;
    ......
    exit(OK);
rollback_and_exit;
    rollback;
    exit(FAIL);
}
```

Box four

Constants

```
R1 = 0; R2 = 0.1; R3 = 0.17;
I1 = 122000; I2 = 36000;
```

Alt. 1

```
IF ((net_income > I1)
AND (net_income < (I1 + I2))) THEN
tax := tax + (net_income - I1) * R2
ELSE IF (net_income >= (I1 + I2)) THEN
tax := tax + I2 * R2 + (net_income - (I1 + I2)) * R3;
```

Alt. 2

```
IF (net_income > I1) THEN
IF (net_income < (I1 + I2)) THEN
tax := tax + (net_income - I1) * R2;
ELSE
tax := tax + I2 * R2 + (net_income - (I1 + I2)) * R3;
```

Alt. 3

```
gov_tax := 0;
IF (net_income > I1) THEN
IF (net_income < (I1 + I2)) THEN
gov_tax = (net_income - I1) * R2;
ELSE
gov_tax = I2 * R2 + (net_income - (I1 + I2)) * R3;
tax := tax + gov_tax;
```

Alt. 4

```
inc1 := 0; inc2 := 0; inc3 := 0;
inc1 := min(I1, net_income);
IF (net_income > I1) THEN
inc2 = min(I2, net_income - I1);
IF (net_income > (I1 + I2)) THEN
inc3 = net_income - I1 - I2;
tax := tax + (inc1 * R1 + inc2 * R2 + inc3 * R3);
```

Alt. 5

```
tax := tax +
min(I1, net_income) * R1 +
min(I2, max(0, net_income - I1)) * R2 +
max(0, (net_income - I1 - I2)) * R3;
```

Alt. 6

```
tax := tax + min(I2, max(0, net_income - I1))
* R2 + max(0, net_income - I1 - I2)) * R3;
```

(The functions min() / max () return the minimum/maximum of two values.)

Box five

Constants

R1 = 0; R2 = 0.1; R3 = 0.17; R4 = 0.2
I1 = 122000; I2 = 36000; I3 = 50000

Alt. 1 - modified

```
IF ((net_income > I1) AND (net_income < (I1 + I2))) THEN
tax := tax + (net_income - I1 ) * R2
ELSE IF ((net_income >= (I1 + I2) AND (net_income < (I1 + I2 + I3))) THEN
tax := tax + I2 * R2 + (net_income - (I1 + I2)) * R3
ELSE IF ((net_income >= (I1 + I2 + I3))) THEN
tax := tax + I2 * R2 + I3 * R3 + (net_income - (I1 + I2 + I3)) * R4;
```

Alt. 5 - modified

```
tax := tax +
min(I1, net_income) * R1 +
min(I2, max(0, net_income - I1)) * R2 +
min(I3, max(0, net_income - I1 - I2)) * R3 +
max(0, net_income - I1 - I2 - I3) * R4;
```

A coaxial notch filter with several notches

BY KNUT N STOKKE

621.315.212

When it is necessary to reduce the interference from a signal at a frequency near the wanted signal, notch filters of the types shown in figure 1 are often used.

In figure 1 is also indicated the amplitude response for such a filter. The length l_1 is about $\lambda/4$ for the interfering frequency, and the amplitude response for this part of the filter is indicated by a dotted curve. The length l_2 is also about $\lambda/4$ for the interfering frequency, and a remnant of the short-circuited $\lambda/2$-resonance is therefore transferred into the resulting response through length X. The resulting response is indicated by the solid curve.

In order to get a sharp filter response, the losses in length l_2, and especially in length X, have to be minimised. The length l_2 must therefore have relatively large dimensions, and this construction will thus be rather expensive. However, it is possible to construct a filter which is simpler and cheaper and may give nearly the same effect.

The length X in figure 1 is part of a transformer between l_1 and l_2, and the current in X is relatively large. It is possible to avoid large currents by using a capacitive coupling instead of an inductive coupling. In figure 2 is shown a short-circuited coaxial cable where the inner conductor is connected to the inner conductor in the signal cable by a coupling condenser.

An open cable can also be used for such a filter, but then the possibilities at low frequencies will be reduced.

If the coupling condenser is short-circuited, the amplitude response will be as shown by the dotted curve in figure 2. The solid curve is the response with the coupling condenser. The result is somewhat different from what is shown in figure 1. Also here we get multiple resonances, but the frequency bands in between are very little influenced by the resonances. In addition we get a resonance at a relatively low frequency.

The resonance at the low frequency is due to the fact that the cable-condenser circuit is capacitive at frequencies near zero. As the frequency increases, the circuit will become more inductive. At a certain frequency, when the length l is less than $\lambda/4$, resonance will occur. We may say that the rejection at zero frequency when C is short-circuited (or $C = \infty$), is moved upwards in frequency depending on the size of the condenser and the characteristic impedance of the cable. This gives an opportunity to construct coaxial stop-filters for frequencies even lower than 30 MHz without using very long cables.

For higher frequencies (100 - 250 MHz) it is often more convenient to use the second resonance where the length of the filter is more than $\lambda/2$ but less than $3\lambda/4$.

Because the resonances are repeated upwards in frequency, we will also get possibilities to construct relatively sharp filters for frequencies up to 1 GHz.

However, in order to take advantage of the possibilities, it is necessary to find a way of calculating the resonances.

The filter consists of a short-circuited cable and a condenser in series. The impedance of a short-circuited cable is:

$$Z_0 = jz_0 tg\left(2\pi \frac{l}{\lambda}\right)$$

where z_0 is the characteristic impedance and l is the length of the cable. The impedance of the filter is then:

$$Z_t = jz_0 tg\left(2\pi \frac{l}{\lambda}\right) + \frac{1}{j\omega C}$$

or:

$$Z_t = jz_0 tg\left(360^\circ \frac{l}{\lambda}\right) - \frac{j}{2\pi fC}$$

The series resonances are when:

$$Z_t = 0 = jz_0 tg\left(360^\circ \frac{l}{\lambda}\right) - \frac{j}{2\pi fC}$$

$$Z_t = jz_0 tg\left(360^\circ \frac{l}{\lambda}\right) - \frac{j}{2\pi fC}$$

Here we want l to be a function of f, and $\lambda = 3 \cdot 10^8 / f\sqrt{\varepsilon_r}$ (m) is used to eliminate λ:

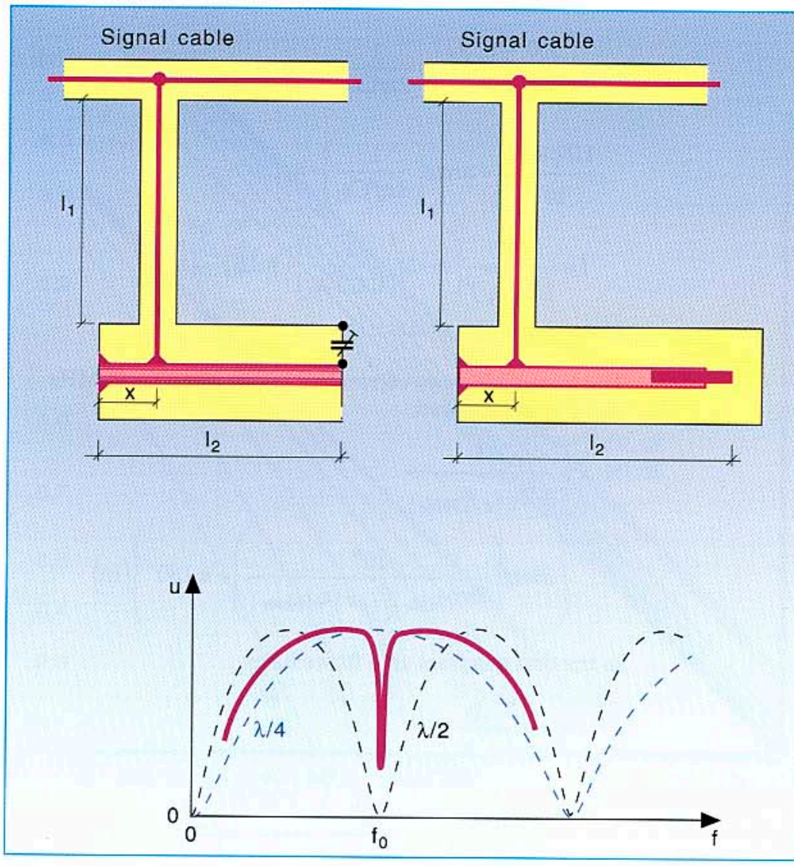


Figure 1 Notch filters

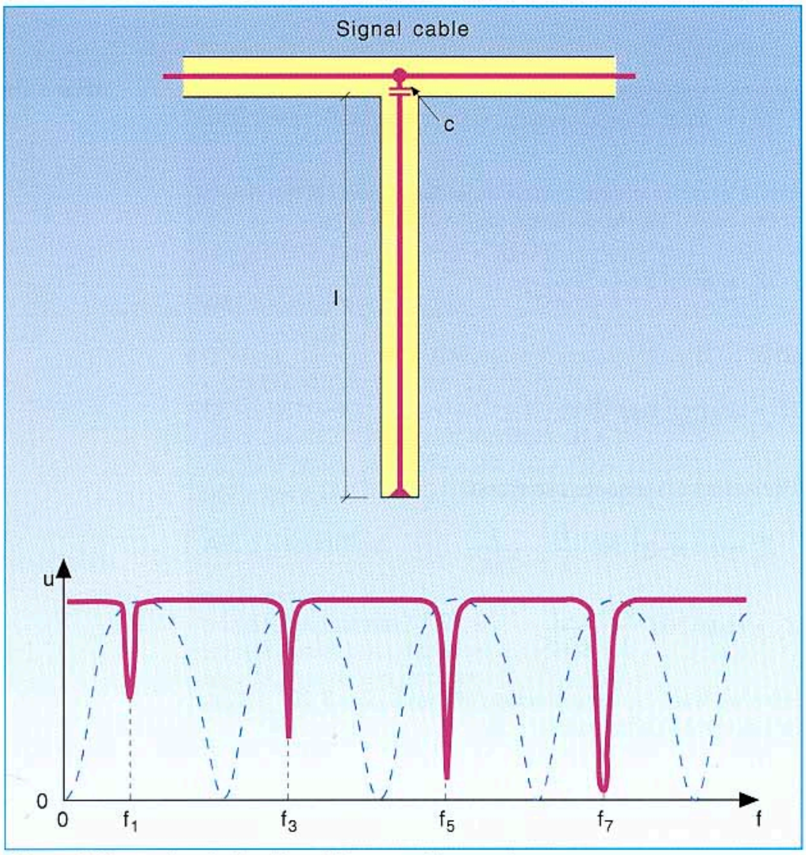


Figure 2 Short-circuited cable with a coupling condenser

$$tg\left(\frac{360^\circ\sqrt{\varepsilon_r}fl}{3.10^8}\right)=\frac{1}{2\pi fCz_0}$$

$$\frac{1.2\sqrt{\varepsilon_r}fl}{10^6}=arctg\left(\frac{1}{2\pi fCz_0}\right)+n.180^\circ$$

$$l=\frac{10^6}{1.2\sqrt{\varepsilon_r}f}\left(arctg\left(\frac{1}{2\pi fCz_0}\right)+n.180^\circ\right)(\mathrm{m})$$

where n is 0 or a positive integer.

The formula is more convenient to use when f is in MHz and C is in picofarad:

$$l=\frac{1}{1.2\sqrt{\varepsilon_r}f_{(MHz)}}\left(arctg\left(\frac{10^6}{2\pi f_{(MHz)}C_{(pF)}z_{0(ohms)}}\right)+n.180^\circ\right)(\mathrm{m})$$

In the first quadrant *(n = 0)* we have:

$$l=\frac{1}{1.2\sqrt{\varepsilon_r}f_{(MHz)}}\left(arctg\left(\frac{10^6}{2\pi f_{(MHz)}C_{(pF)}z_{0(ohms)}}\right)\right)(\mathrm{m})$$

When C varies towards zero, l approaches the limit:

$$\lambda/4=\frac{75}{f_{(MHz)}\sqrt{\varepsilon_r}}(\mathrm{m})$$

The other limit is when $C=\infty$, i.e. $l=0$.

When the values from the first quadrant is used, we will get a reasonable length for the filter even at relatively low frequencies.

In the third quadrant *(n = 1)* we have:

$$l=\frac{1}{1.2\sqrt{\varepsilon_r}f_{(MHz)}}\left(arctg\left(\frac{10^6}{2\pi f_{(MHz)}C_{(pF)}z_{0(ohms)}}\right)+180^\circ\right)(\mathrm{m})$$

The limits are then:

$$l=\lambda/2=\frac{150}{f_{(MHz)}\sqrt{\varepsilon_r}}(\mathrm{m})$$

and:

$$l=3\lambda/4=\frac{225}{f_{(MHz)}\sqrt{\varepsilon_r}}(\mathrm{m})$$

In figures 3 - 6 curves in the first and the third quadrant for some values of C are given. The curves are calculated for 50 ohm and 75 ohm cable with relative permittivity $\varepsilon_r = 1$ and $\varepsilon_r = 2.3$.

We can see that in the third quadrant there are only small differences between the curves for 50 ohm cable and 75 ohm cable when ε_r is the same. The differences become more important in the first quadrant when large condensers are used. However, if we want to construct a filter of 60 ohm cable, we may get sufficient accuracy by using the mean value between the 50 ohm curves and the 75 ohm curves.

As can be seen from figure 2, the resonances are repeated upwards in the frequency bands. If the higher resonances are used, the filter may be constructed for frequencies up to 1 GHz. When such a filter is to be constructed, it is not necessary to draw curves for all the quadrants where the resonances occur. If we look at the equation for the length l of the filter and the curves in the figures 3 - 6, we can see that with small coupling condensers (C < 5 pF), the ratio between the resonant frequencies in the first and the third quadrant is about 1 : 3. If we look at higher frequencies, the resonances are repeated every 180°, and the ratios between the resonant frequencies will therefore be the odd numbers:

1 : 3 : 5 : 7 : 9 : ... .

When we want to construct a relatively sharp filter for higher frequencies, the condenser has to be small. A filter which is calculated for 50 MHz in the first quadrant, will also have resonances at 150 MHz, 250 MHz, 350 MHz, and so on. If we for instance want to construct a stop filter for

500 MHz, we may as a starting point use a filter which in the first quadrant has series resonance at 500/9 MHz = 55.56 MHz or at 500/7 MHz = 71.43 MHz.

When we use higher resonances, the corresponding voltage distributions will be as indicated in figure 7. If we then will insert a variable condenser in order to tune the filter, we must therefore take into consideration the voltage distribution in the standing wave. In addition the filter must be shortened because of the introduction of an extra capacitance. In most cases the shortening is made in such a way that the wanted resonant frequency is in the middle of the tuning range.

If the tuning condenser is placed at a voltage maximum, we get the maximum tuning range. But also the mechanical instabilities in the condenser will then have maximum influence. It may therefore be convenient in some cases to place the condenser at a lower voltage, but then also the tuning range will be reduced.

In order to calculate the tuning range, it is necessary to find the capacity per unit length of the cable. The capacitance of a coaxial cable is:

$$C_c = \frac{2\pi\varepsilon_0\varepsilon_r l}{ln\frac{d_2}{d_1}}$$

where d_2 is the inner diameter of the outer conductor and d_1 is the diameter of the inner conductor. l is the length of the cable.

The characteristic impedance of a coaxial cable is:

$$z_0 = \frac{60}{\sqrt{\varepsilon_r}} ln\frac{d_2}{d_1}\ ):\ ln\frac{d_2}{d_1} = \frac{z_0\sqrt{\varepsilon_r}}{60}$$

and when this value and $\varepsilon_0 = \frac{10^{-9}}{36\pi}\left(\mathrm{F/m}\right)$

are introduced into the equation for C_c, we have:

$$C_c = \frac{100\sqrt{\varepsilon_r}}{3z_0}\ \left(\mathrm{pF/cm}\right)$$

For 50 ohm cable with $\varepsilon_r = 1$, the result is:

$C_{50(1)} \approx 0.67$ pF/cm

and for 50 ohm cable with $\varepsilon_r = 2.3$:

$C_{50(2.3)} \approx 1$ pF/cm.

For 75 ohm cable:

$C_{75(1)} \approx 0.44$ pF/cm

$C_{75(2.3)} \approx 0.67$ pF/cm.

The inductance of the cable is (20). However, when we want to calculate the tuning range, it is sufficient to use the capacitance. If for example the tuning condenser has a maximum capacitance of 12 pF and is placed in a voltage maximum for a 50 ohm cable with $\varepsilon_r = 2.3$, we may say that

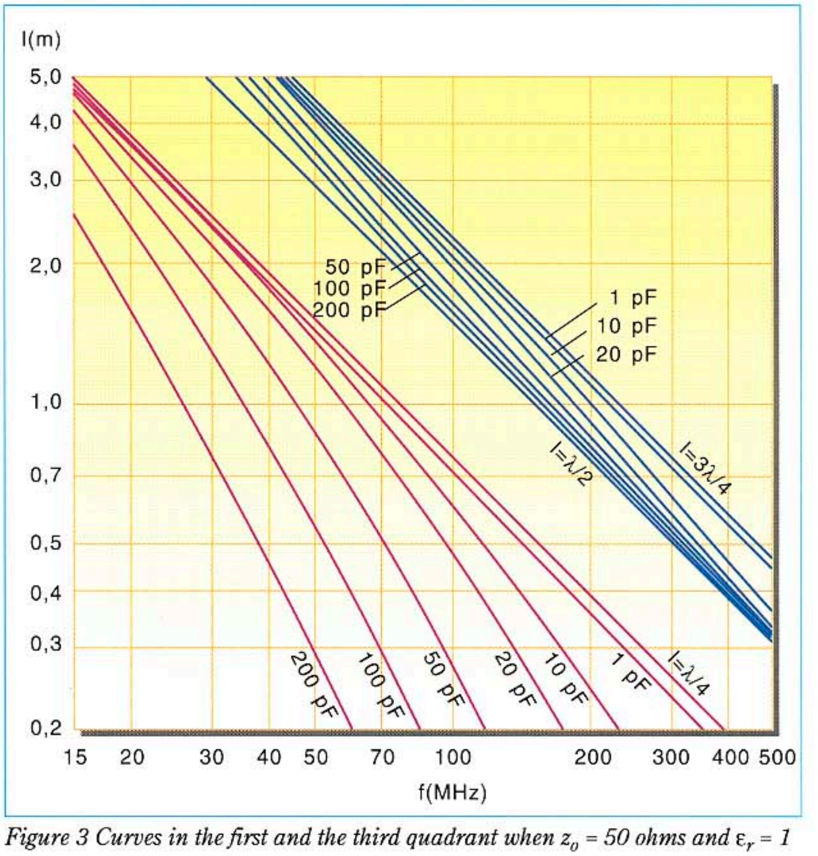


Figure 3 Curves in the first and the third quadrant when z_0 = 50 ohms and ε_r = 1

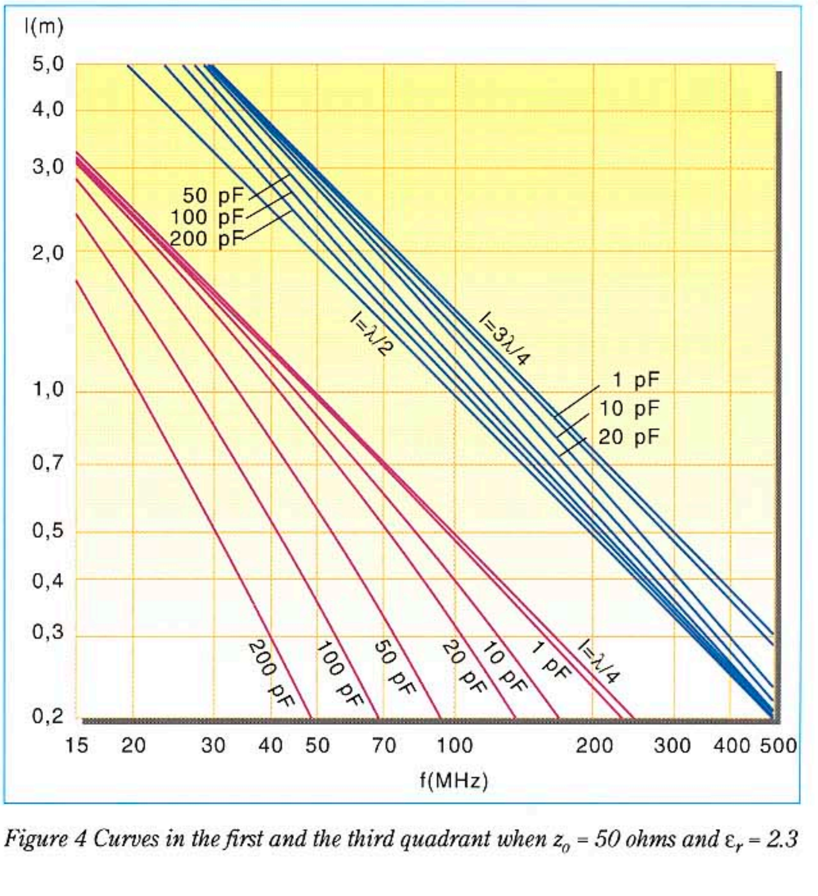


Figure 4 Curves in the first and the third quadrant when z_0 = 50 ohms and ε_r = 2.3

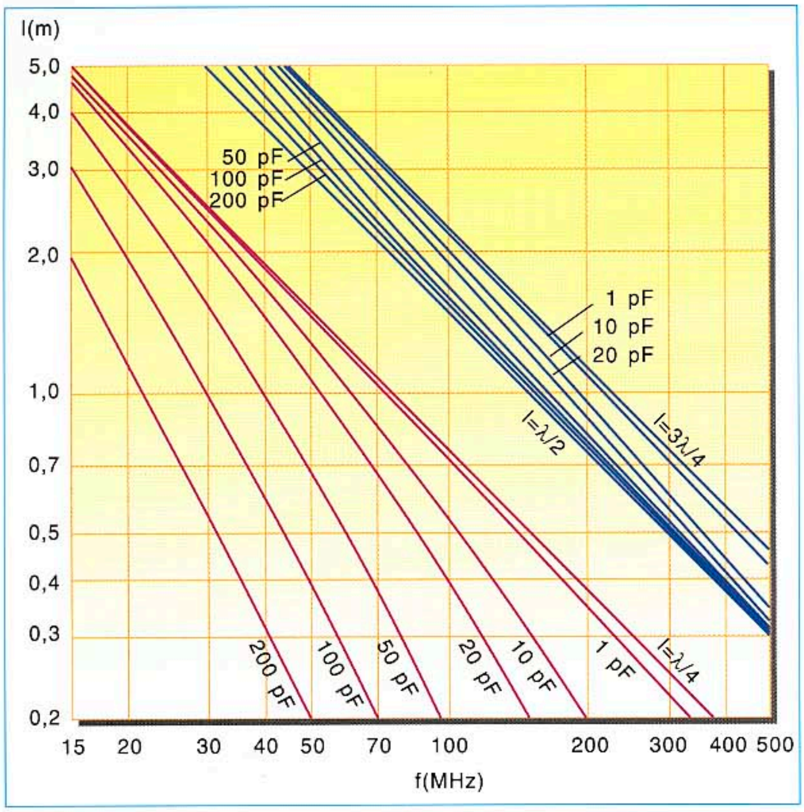


Figure 5 Curves in the first and the third quadrant when z_0 = 75 ohms and ε_r = 1

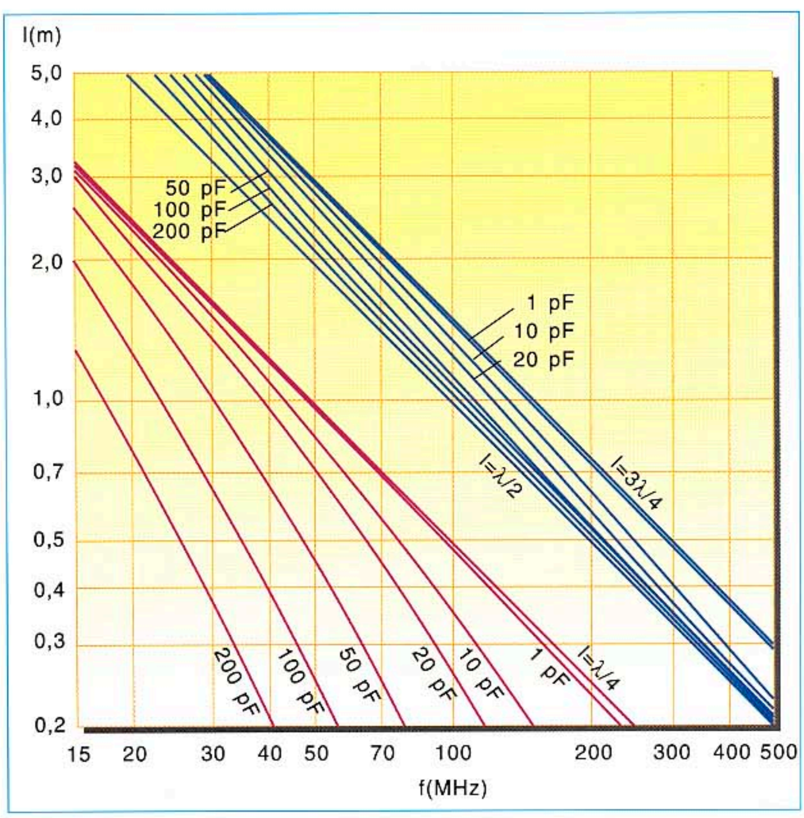


Figure 6 Curves in the first and the third quadrant when z_0 = 75 ohms and ε_r = 2.3

we can shorten the filter about 12 cm and manage to tune the resonance back to the original frequency.

It may be convenient to have the actual rejection frequency in the middle of the tuning range. When the tuning range of the condenser is 2 - 12 pF, we may write the capacitance as (7 ± 5) pF. If for example the filter is made of 50 ohm cable with $\varepsilon_r = 2.3$, and we use the resonance in the third quadrant, we may shorten the filter about 7 cm and get a tuning range around the wanted resonant frequency when the condenser is placed at a voltage maximum.

We have to remember that the absolute maximum tuning range we can get when a tuning condenser is placed at a voltage maximum, is when we tune from full voltage to nearly 0 voltage (short circuit or large condenser). The method of calculating the tuning range described here, is therefore valid only when the shortening is small compared with the wavelength.

If the tuning condenser is placed outside the voltage maxima, this must be taken into account when calculating the tuning range. The voltage distribution in a standing wave is sinusoidal, and this may be used to find the voltage over the tuning condenser.

We take as an example a filter for 187.75 MHz made of 50 ohm cable with $\varepsilon_r = 2.3$, and with a 3 pF coupling condenser. When we use the values in the third quadrant, the length of the filter is found to be *76 cm*. The wavelength in the filter cable is 105 cm. If a tuning condenser of 2 - 12 pF is placed 15 cm from the short-circuited end of the cable, the voltage over the condenser will be reduced to:

$$\sin\left(\frac{15}{105}\cdot 360^\circ\right) = 0.78$$

We may then say that the influence of the condenser will be approximately the same as if it was placed in a voltage maximum with the value:

$$0.78(7 \pm 5)\ \text{pF} \approx 5.5 \pm 4\ \text{pF}.$$

For 50 ohm cable with $\varepsilon_r = 2.3$ we have about 1 pF/cm, and the filter cable may be shortened about 5.5 cm, that is to about 70.5 cm. When the condenser varies ±4 pF, this will compensate for a length variation of the cable from about 80 cm to about 72 cm. From the curves in figure 4 we then get a tuning range of about:

178 - 200 MHz.

Concerning the attenuation which may be achieved at the resonant frequencies, this is dependent on the losses in the cable and in the coupling condenser. Because of relatively high voltage over the coupling condenser, this condenser must have low losses, especially when we use the filter for higher frequencies. Good ceramic condensers or other condensers with high Q may be used.

The losses in the cable is dependent on dimensions, materials and also on the characteristic impedance. The characteristic impedance which gives minimum losses may be calculated by using figure 8.

Attenuation (in Nepers) in a cable is:

$$\alpha = \frac{R}{2z_0}$$

where R is the resistance in the inner and outer conductor of the length l of the cable.

If the resistivity for the inner conductor is ρ and the resistivity for the outer conductor is $v \cdot \rho$, the high frequency resistance of the coaxial cable is:

$$R = \frac{\rho l}{F_1} + \frac{v\rho l}{F_2} = \frac{\rho l}{\pi\delta}\left(\frac{1}{d_1} + \frac{v}{d_2}\right)$$

F_1 and F_2 are the effective conducting areas and δ is the penetration depth. For the actual frequencies the penetration depth is very small compared with the dimensions of the cable.

The attenuation because of the resistance in a coaxial cable is then:

$$\alpha = \frac{\frac{\rho l}{\pi\delta}\left(\frac{1}{d_1} + \frac{v}{d_2}\right)}{2 \cdot \frac{60}{\sqrt{\varepsilon_r}} ln \frac{d_2}{d_1}} = \frac{\rho l \sqrt{\varepsilon_r}}{120\pi\delta} \cdot \frac{1}{d_2}\left(\frac{\frac{d_2}{d_1} + v}{ln \frac{d_2}{d_1}}\right)$$

We may in this calculation consider d_2 to be constant and d_1, that is, d_2/d_1 as variable:

$$\alpha = k' \frac{\frac{d_2}{d_1} + v}{ln \frac{d_2}{d_1}} = k' \frac{x + v}{ln x}$$

$\frac{d\alpha}{dx} = 0$, in this case a minimum value, gives $ln x = 1 + \frac{v}{x}$.

If the inner and outer conductor are of the same material, we have:

$$ln x = 1 + \frac{1}{x}$$

This equation may for instance be solved graphically, and the solution is that the minimum value is when $d_2/d_1 = 3.59$ or $ln(d_2/d_1) = 1.278$.

If $\varepsilon_r = 1$ (dry air, vacuum), the characteristic impedance which gives minimum losses is:

$z_{0(min)1} = 76.7$ ohms.

When $\varepsilon_r = 2.3$ (solid polyethylene), we have:

$z_{0(min)2.3} = 50.6$ ohms.

Some types of cables have copper in the inner conductor and aluminium in the outer conductor. The ratio between the resistivities for aluminium and copper is 2.828/1.724 = 1.64. The solution of the equation gives $ln(d_2/d_1) = 1.4$, and when $\varepsilon_r = 1$, we have:

$z_{0(min)1} = 84$ ohms,

and when $\varepsilon_r = 2.3$:

$z_{0(min)2.3} = 55.4$ ohms.

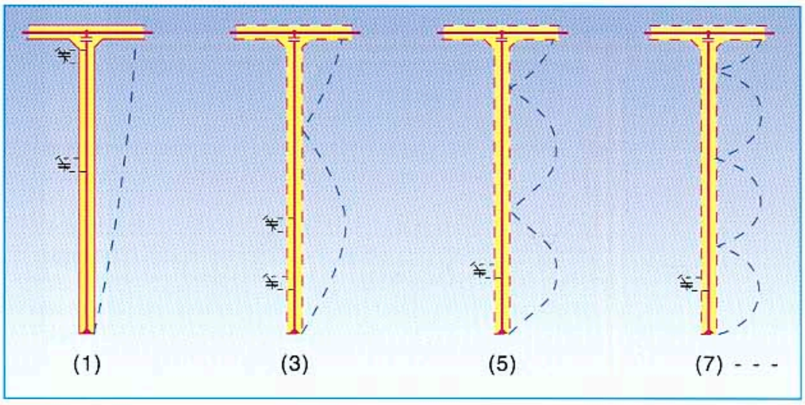


Figure 7 Voltage distribution in the filter at diffrent resonant frequencies

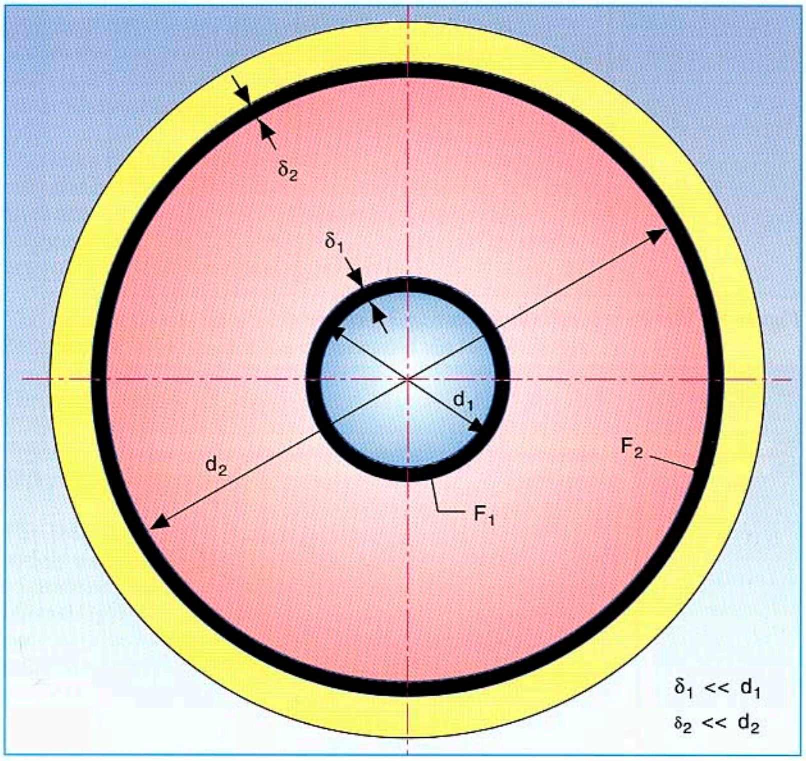


Figure 8 Diameters and penetration depth in a coaxial cable

If we have the same material in the inner and the outer conductor and $\varepsilon_r = 1$, 75 ohm cable should give a good result. If $\varepsilon_r = 2.3$, 50 ohm cable is a convenient characteristic impedance.

When the length of the filter is in the order of 1 or 2 metres and $\varepsilon_r = 2.3$, there are only small differences if we use 75 ohm cable or 50 ohm cable in the filter. The losses are more dependent on the materials and the dimensions. Large dimensions and good conducting materials give small losses and therefore good efficiency to the filter. Silver plated materials are effective because the penetration depth is so small at the actual frequencies that most of the current will go in the silver layer.

If we want to use two filters in order to get more attenuation, the length of the signal cable between them should be $\lambda/4$ for the rejection frequency. With $\lambda/4$ between the filters

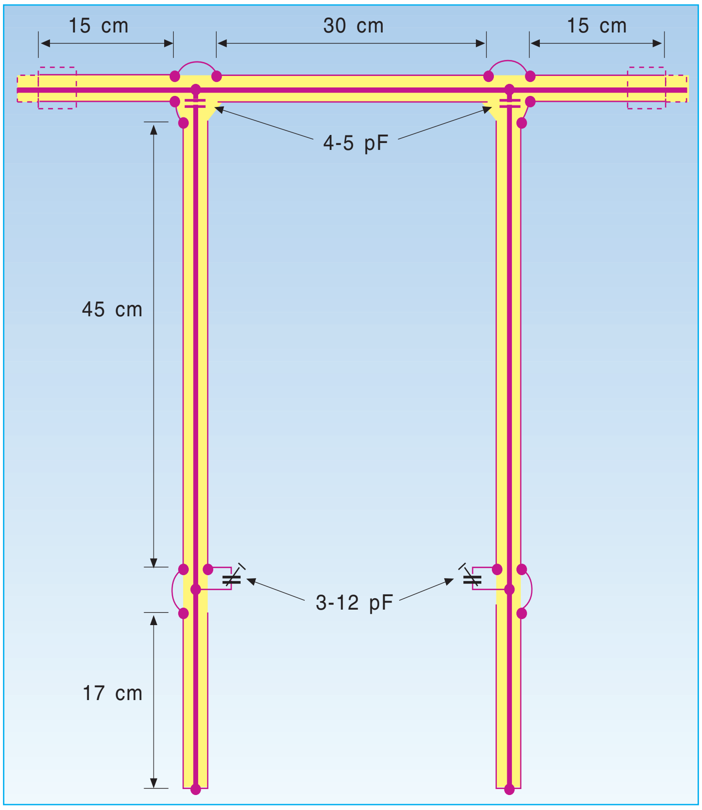


Figure 9 Filter to reduce interference in TV channel 5

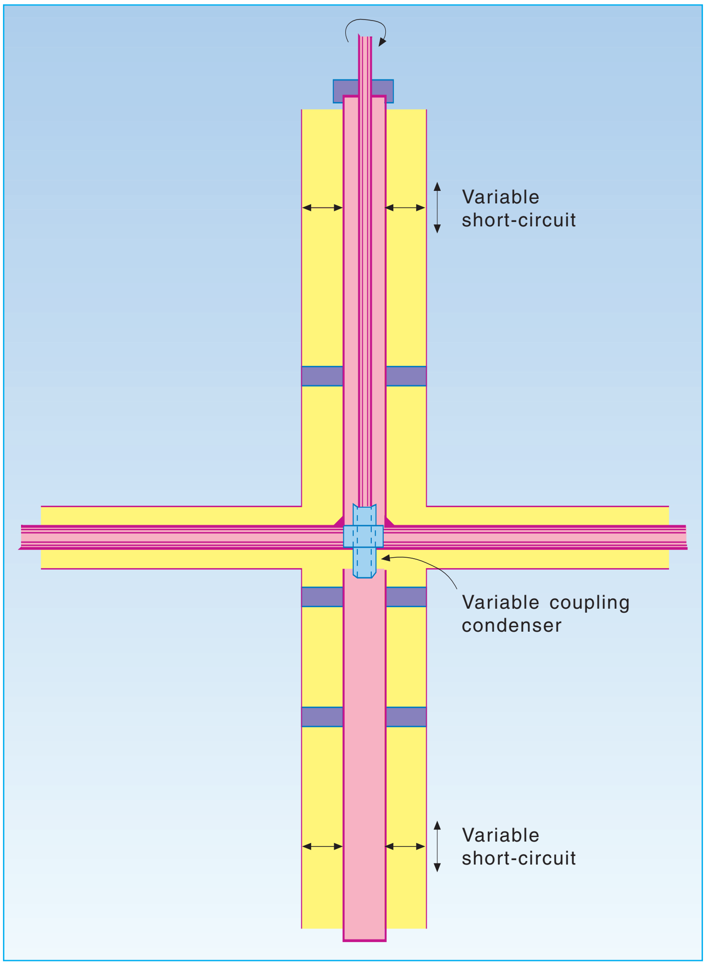


Figure 11

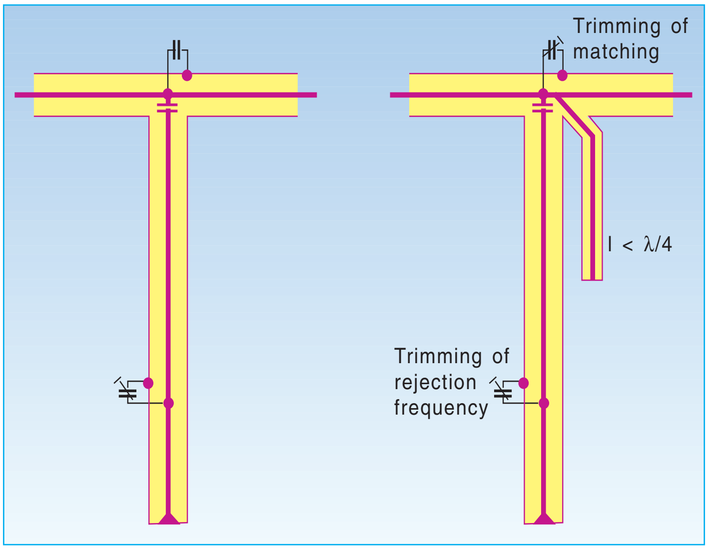


Figure 10 Matching of the filter

between them is at a minimum, and the attenuation in dB is doubled. In practice we may get more than the double attenuation because the resulting filter attenuation curve will be relatively broadened.

In figure 9 is shown the principle of a double filter to protect TV receivers in channel 5 against interference from transmissions on frequencies just below channel 5. This filter is made of ordinary 75 ohm cable for receiving installations. The attenuation at for instance 171 MHz is more than 30 dB, and the loss at the channel 5 picture carrier (175.25 MHz) is only 1.5 dB.

Concerning the matching, this filter normally gives satisfactory matching for the wanted signal. However, when the filter is used very near the wanted signal, it may be necessary to have better matching for the wanted signal.

If we want to have good matching conditions on frequencies just over the rejection frequency, the matched frequency band has to be tuned to lower frequencies. This can simply be done by a condenser parallel over the signal cable, as shown to the left in figure 10. For instance if we want to use the mentioned filter for 187.75 MHz to protect TV channel 7 against interference from the sound signal in TV channel 6, a condenser of 2 - 5 pF may give satisfactory matching for TV channel 7.

If, on the other hand, we want better matching just below the rejection frequency, the matched frequency band has to be tuned to higher frequencies. We must then have more inductance over the signal cable. This may be obtained by using a short-circuited cable shorter than $\lambda/4$ parallel over the signal cable, as shown to the right in figure 10. The losses in this short-circuited cable have little influence on the filter, and may therefore be made of thin cable. If in addition a trimming condenser is used, we can adjust the matched frequency band. However, when such extra components are used, the amplitude response for other frequencies is changed. The amplitude response will therefore not be so simple as shown in figure 2.

NTR's high speed testbeds and multimedia activities

BY ØIVIND KURE, INGVILD SORTEBERG, TERJE ORMHAUG AND RUNE FLØISBONN

654.1:681.3

Abstract

This article gives an overview of the high speed testbeds and multimedia research activities within the Norwegian Telecom. The internal testbeds and the multimedia activities have been carried out under the A-lab and B-lab research programs.

1 Introduction

A-lab, application lab, and B-lab, broadband lab, are two of the major research programs at NTR. A-lab includes activities in multimedia application development, high speed protocols, interworking between private and public networks and interface cards for ATM end systems. B-lab is the development of an experimental ATM switch and its control system. The switch is specially designed to be used as a basis for ATM traffic studies.

Currently, there are three different high speed testbeds, 1) Supernet, a 34 Mbit/s Wide Area Network (WAN), 2) a 140 Mbit/s Local Area Network based on ATM (LATM), and 3) an experimental ATM testbed. Both 2 and 3 are located at the Norwegian Telecom Research lab. In conjunction with these testbeds, NTR has undertaken several research projects in high speed protocols, end systems' ATM interface cards, and multimedia applications like multimedia post systems, multimedia conferencing, and distance education.

The main ideas behind the article are the description of and the philosophy behind the testbeds and their interconnection, the activities in high speed protocol, the multimedia applications, in particularly the distance education application, and a summary of some of the still unsolved problems in multimedia application development. The article also includes a summary of some of NTR's future research plans within these areas.

Section two outlines the testbeds, their architectures, objectives and research goals. The third section describes the interconnection of the testbeds. Section four outlines our work on high speed protocols for ATM networks. Section five describes NTR's participation in end system ATM interface card development. An integral part of the testbeds is the demonstration applications. In our testbeds, distance education is one of the more unique ones. An outline of this prototype application is therefore included in section six. This section also summarises some of the problems experienced when developing multimedia applications.

2 Description of testbeds

The testbeds reflect Norwegian Telecom's national and international co-operation. The WAN testbed, Supernet, is the result of a joint project between the Norwegian Telecom and the four Norwegian universities. The ATM LAN testbed is highly correlated with the work in RACE project 2060, in particular the development of high speed transport and network protocols. The experimental ATM testbed is the result of our participation in RACE projects 1022 and 1083.

One of the unique aspects of the testbeds is the degree of interconnection between them. The WAN and the LAN testbeds are connected, while the experimental ATM testbed will be connected during 1993.

None of the testbeds are under the same administrative management. Each is organised and managed by separate teams. However, by emphasising interconnection at various levels, the three testbeds can utilise each other, and thereby expand their individual usefulness.

2.1 Supernet

Supernet is a joint project between Norwegian Telecom and the four Norwegian universities. It is the largest testbed geographically with a diameter of more than 1800 km. It covers the universities' campuses in Oslo, Bergen, Trondheim and Tromsø and the Norwegian Telecom Research lab at Kjeller (figure 1).

The links have 34 Mbit/s bandwidth. These are multiplexed onto the backup lines in the telephone network. The backup lines are only available for Supernet as long as there are no outages in the telephone network. The availability of the network can therefore not be guaranteed.

The architecture is similar to the one reported in (1). The 34 Mbit/s PDH links are interconnected via Network System's routers. There is one router at each of the five sites. The routers have in addition to the 34 Mbit/s interface, one FDDI interface and up to four Ethernet interfaces. The Ethernet interfaces are mainly used for management and experiments.

The DoD IP protocol is used in Supernet. This protocol is widely used at the universities, and routers with the required interfaces were available. Supernet is the first IP service offered by Norwegian Telecom. It is used to gain experience with IP networks and as a basis for working on unsolved problems related to management and accounting of IP services.

Norwegian Telecom is responsible for management of the links and the FDDI rings connected directly to the routers. Manage-

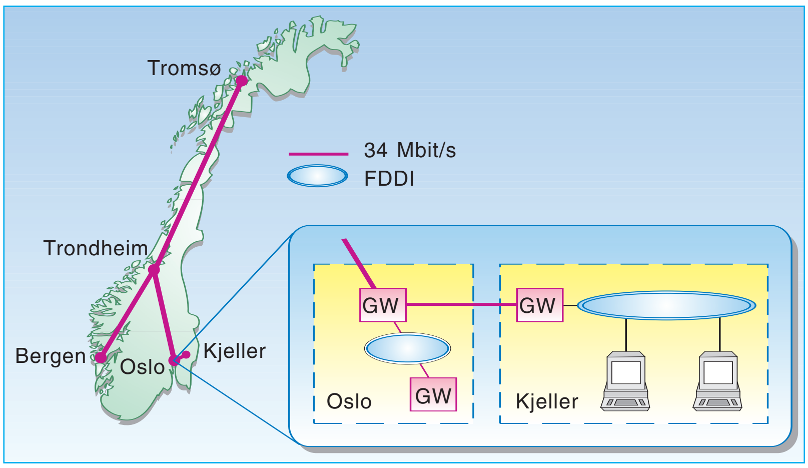


Figure 1 Topology of Supernet

ment is done by use of SNMP from a central location. SNMP information is exchanged between Norwegian Telecom and the universities. The Norwegian Telecom Research (NTR) connects their workstations directly to the same FDDI ring as the router. The universities typically use another architecture. They isolate their networks by connecting only one router to the Norwegian Telecom's FDDI ring. The other side of such a router is then connected to the router backbone within the university. This firewall architecture increases the cost, but provides a clear separation between different management and administrative domains.

On top of this framework there are several research projects for high speed applications. Two of these are distance education using electronic whiteboard and video conferencing. The former is described in section 6.2.5. An other use of Supernet is to provide access to the supercomputers at different universities' sites.

Supernet is not a static network and new technologies will be used to enhance it. In the first phase this might be an evolution towards a WAN ATM network based on semi-permanent virtual channels. This implies that management routines are used for connection establishment and release. In the future, extensions of Supernet might be to include other subscribers. In particular other Norwegian research institutes and the oil industry have expressed great interest in joining.

2.2 NTR testbeds

The Norwegian Telecom Research maintains two internal communication testbeds in addition to the FDDI ring connected to Supernet. The FDDI ring connects four workstations and the Supernet router. The other internal testbeds are a local area network based on ATM (LATM) and an experimental ATM testbed. Each of these are described in separate subsections.

Two of them, the FDDI ring and the LATM network, are organised as high speed local area network testbeds. A workstation is used as an IP router between the ATM network and the FDDI. This will be augmented with an XTP router developed within the RACE II project 2060, CIO.

2.2.1 FDDI

The FDDI ring started out as a testbed for high speed protocols. It has now evolved to include two FDDI rings one used as a backbone and the other as a testbed. Our lab has more than 250 workstations connected to a local area network. These are evenly distributed in four interconnected buildings. The local area network could no longer support such a large number of workstations. It was therefore replaced with an FDDI backbone. The backbone connects two servers plus four hubs installed in the four corners of our building complex. The FDDI test network is connected to the backbone via the hubs. The Supernet router is connected to the testbed FDDI ring. Individual workstations with FDDI connections are either connected directly to the test ring or through bridges and routers to the backbone network. Currently, only four workstations are connected directly to the FDDI test network.

2.2.2 Local area ATM network

The LATM testbed has three goals, 1) to gain operational experience with local ATM network, 2) to demonstrate ATM capabilities, and 3) to act as a platform for research on data communication over ATM. This is reflected in an architecture with a large degree of connectivity. This allows rapid changes in the configuration of the testbed.

The testbed is built around an 8 x 8 ATM switch for LAN from Fore Systems, Inc., (4). ATM for wide area networks and local area networks differ in the following aspects: 1) switch design, 2) signalling protocols, 3) support for ATM Adaptation Layer (AAL), and 4) policing (5). The local ATM switch design is typically simpler since a switch for LATM does not need to be super reliable and with only limited ability to scale. The goal of the project is not switch design so these particular aspects do not impose any restrictions.

Currently, five workstations are connected to the local ATM network. The network only runs IP, but a planned extension to the interface drivers will allow us to run XTP as well. The network will be used to test out applications and protocols developed in the CIO RACE project.

2.2.3 Experimental ATM testbed

The last testbed is an ATM testbed based on NTR's B-lab project and participation in RACE project 1022. One of the goals is to investigate ATM performance for different traffic patterns. This testbed is developed as a workshop for different broadband and control related activities at NTR. The transfer principle of the testbed is 155 Mbit/s ATM. It will consist of a private-network-like switch B-NT2, a public-network-like switch B-LEX-1, a traffic generator MULTIGEN, control functions for set-up of point to point and point to multipoint connections, and console functions for set-up and execution of test scenarios and traffic experiments. The first version of the testbed will be ready early 1993.

One main objective for developing the testbed has been to establish a platform for traffic experiments and verification of ATM functions. The purpose of the MULTIGEN is to make it possible to load the network with "realistic" background traffic. Requirements for this function have been to be able to load the network from a minimum number of physical inlets (8), to program the traffic profile on each inlet to represent a selectable number of sources (from a few to several hundreds), and let each represent a specific source type like data, video, telephony, etc. The programmable traffic model is the same as the one being implemented for the foreground traffic generator in the RACE project 1083, PARASOL.

The B-LEX-1 is a switch with specific test facilities. It consists of three 8 x 8 switching blocks that may be rearranged for different test scenarios. The switching blocks have output buffers with programmable size. A mechanism for cell-loss priority is implemented, it may be enabled or disabled. The cells are timestamped on the inlets, and if they are lost in the outlet buffers, they may be routed to overflow buffers where they can be picked up and examined.

The B-NT2 has been developed as a pilot to investigate possible access networks for ATM. A ring topology with a DQDB-like

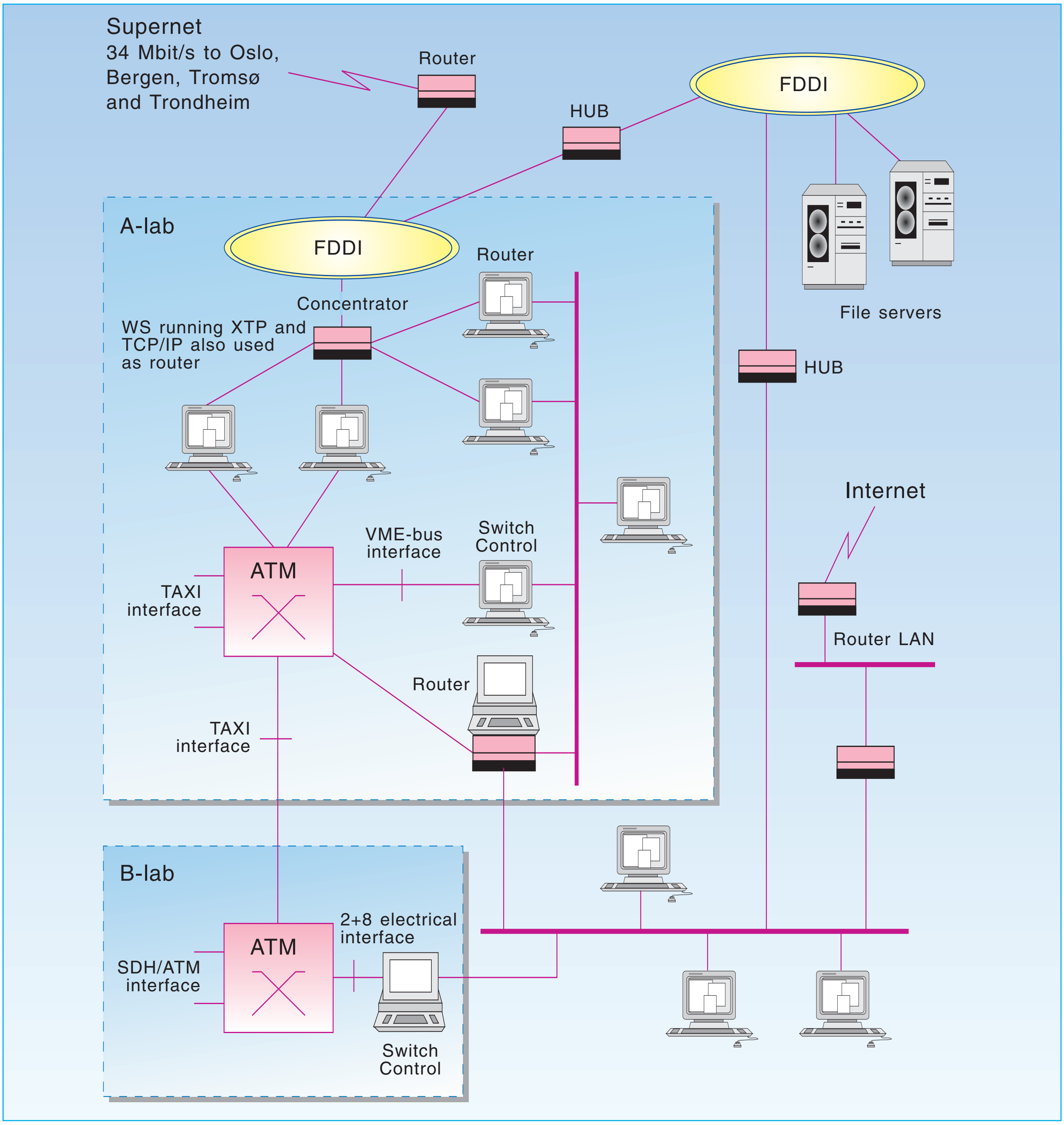


Figure 2 Architecture of the NTR testbeds

access protocol was chosen for this switch, which may represent an NT2 for private customers or small business customers. In the switch external traffic may be given priority over internal traffic. External connections must be set up with call-control functions as in the B-LEX-1, but with the priority mechanism it is possible then to let the internal traffic utilise the free capacity for connectionless communication.

Another main objective is to be a testbed for IN (Intelligent Network) supported B-ISDN connection control functionality. By implementing a connection control system for the pilot network one may verify concepts and ideas that are developed for the IN architecture, and explore interactions between services and the ATM network.

The system control is equipment specific, and for B-LEX-1 it has been realised as a trial for using direct translation from specification language (SDL) to a programming language (OCCAM).

Beyond the first version of the experimental testbed we plan to establish interconnection with other networks and make interfaces for connection of workstations and other end systems.

3 Testbed interconnection

All three testbeds are either connected or will be connected. Our in-house FDDI network is connected to the 34 Mbit/s Supernet. Using a workstation as router, the local ATM network is connected to the FDDI network. The two networks are also connected to the Internet. For our in-house experimental ATM switch, line interface to the local ATM network will be developed. ATM traffic from the local environment can then be routed through the experimental wide area switch. Workstations are connected directly to both the FDDI network and the local ATM network. In a discussion of interconnection, three aspects should be considered, the cabling, the physical connection and the protocol stack.

3.1 Cabling

The in-house testbeds are based on structured cabling with multi-mode fibre. AT&T strategic points optical patch panels are installed. There are two distinct sets of patch panels, one located to serve as a backbone, and one located to interconnect workstations placed on the same floor in one of the wings of our laboratory.

The structured cabling has two distinct advantages, 1) it is simple to restructure the network, and 2) the cabling can be used for both star and ring based technologies. The obvious disadvantage is the increase in cost. Our in-house testbed consists of both LATM and FDDI. By basing it on structured multi-mode optical fibre we have gained a substantial flexibility in the configuration of the testbed.

3.2 Physical connection

The physical interconnection is only an issue for platforms running the same MAC layer, or in our case those running ATM. The experimental ATM switch has an electrical interface consisting of 8 data lines in parallel, a cell line and a byte line. The switch can handle bandwidths between 140 Mbit/s and 155 Mbit/s.

The local ATM uses the TAXI chipset from AMD for physical framing. The bits are transferred over the FDDI physical interface running at 175 MHz or 140 Mbit/s. The framing conforms to the one proposed in the ATM Forum's UNI Specification, (6).

An addition to the experimental ATM switch, to include a physical interface based on the same technology as the local ATM switch, is planned. The only problem is the handling of idle cells, and the treatment of malformed cells. The difference in line speed does not pose a problem since our switch design can handle lower speed than 155 Mbit/s. The interconnection will be available in 1993. This will allow us to interconnect the two testbeds, and thereby be able to run real data communication traffic into the experimental switch and use it in our traffic experiments.

4 Protocol stacks

The IP protocol is used as the main network protocol. The choice of protocol was a pragmatic one, there is a large number of equipment and applications available for the Internet IP protocol stack. Having limited development resources, we depend on buying equipment for most of our experiments. The availability of equipment was therefore important.

The availability of applications and equipment were not the only reasons to choose the Internet protocol stack. In addition, IP has a growing share of the private network market. Interconnection of private networks will be an important initial market segment for ATM. Any traffic experiments should therefore have the option of using workloads consisting of IP applications to simulate the initial workload for ATM.

All workstations connected to the testbeds are assigned IP numbers. More important all network addresses have connected status in the Internet. All elements in the testbeds can therefore be addressed through the various international research networks. This simplifies exchange of results and testing of applications and procedures, particularly in RACE project 2060.

Through this set-up two valuable results are achieved. By connecting the various testbeds with IP routers, all machines are accessible from any location. This gives us an enormous freedom in the application experiments. In addition, we believe strongly in the importance of real world experience. The workstations in the testbeds are used as ordinary stations in the daily life of our labs. In other words, the testbeds are integrated directly in the operational internal network. This clearly creates problems, but it gives us invaluable experience and added insights into the problems.

4.1 XTP (Xpress Transfer Protocol)

XTP is a new lightweight transfer protocol designed for high speed networks, (3). The protocol includes both layer 3 and 4, and the design is optimised for implementation in VLSI. In our testbed, a software version of XTP, the XTP Kernel Reference Model delivered by Protocol Engine Inc., is used.

XTP uses a connection oriented approach based on a virtual channel (path) establishment. A set of Quality of Service (QoS) parameters are associated with the channel indicating the user's requested burst and rate values. XTP provides mechanisms for resource reservation, but these are currently not implemented. Across the channel, XTP offers connection oriented, connectionless (both acknowledged and unacknowledged), transaction and isochronous services. The protocol mechanisms in XTP are flexible and well suited to provide transport services for multimedia applications.

XTP is well suited to utilise the services offered by ATM, particularly the channel concept of XTP makes it possible to associate an ATM VC/VP directly with an XTP virtual channel. This enables a tighter coupling of the QoS requested by the user and the QoS offered by ATM. Specification of XTP over ATM will be done as part of the RACE project 2060. A specification of XTP over ATM includes in addition to the protocol encapsulation problem, issues like resource reservation and signalling. In B-ISDN, outband signalling is used, while XTP uses inband signalling. These two concepts must be coupled in order to allow a tight connection of the virtual channels in XTP and ATM. The user can signal the required service across the transport protocol service interface, and these parameters are mapped to B-ISDN signalling procedures. Network resources are not the only ones needed to be reserved. To guarantee a particular QoS, end system resources like buffers and CPU time slices must be reserved.

Currently the XTP KRM implementation only runs over Ethernet and FDDI networks, and we do not have an XTP router. In RACE project 2060, how to run XTP over ATM and an XTP router will be specified and implemented. The router will be implemented on a workstation. Hence, work done on signalling and resource reservation for end systems will be used in the development of the XTP router. Another challenge in the router development is to explore how a service guarantee can be maintained across networks based on different technologies and offering different services. The XTP router will be integrated into the NTR testbed, giving us a unique possibility to experiment with new network and transport services.

4.2 Future plans

The IP protocol stack is not the ideal protocol stack for ATM testbeds. Applications running over IP can only utilise the high bandwidth offered by ATM. Rough calculations on the processing overhead in IP indicate an upper limit around 100 Mbit/s (2). From a bandwidth view point, IP is therefore sufficient to utilise ATM's potential.

However, ATM also offers a channel concept, with the potential to offer different Quality of Service (QoS) for a substantial number of virtual channels. The IP protocol is badly suited to take advantage of this concept. Although the transport protocol, TCP, has a QoS field, it is not used. This is also reflected in the commercial implementations of IP over ATM, where all IP traffic with the same source and destination addresses is multiplexed on the same virtual channel.

In our opinion, successful use of ATM may depend on the development of transport and network protocols that can take advantage of the concept of channels with individual Quality of Service. An effort to adapt and implement such protocols is therefore part of one of the testbeds. XTP is an example of a protocol that fits this descriptions.

5 ATM interface card development

The testbeds have so far relied on commercial available end systems. These interfaces represent a bottleneck for the transfer of data. The testbed activity is augmented by an effort to develop an ATM interface for multiprocessor end systems. The term multiprocessor is used loosely, and it encompasses end systems with separate video and/or protocol processors. The main idea is to avoid the bottleneck represented by the backplane bus. In addition, the interface should be able to route ATM virtual channels directly to the destination processor. With this architecture, one can avoid interference between a video stream and the more "regular" data communication stream in the backplane and in the CPU.

An ATM interface for the Scalable Coherent Interface (SCI) is in the design phase. SCI is an IEEE standard for backplane bus. It offers 1 Gigabyte/s bandwidth on a ring. Data is transported in chunks of 64 bytes. SCI also offers memory coherency between different nodes.

The interface will be a simplistic one. It will map between memory address and virtual channel in ATM. There will also be some support for rate control and CRC checking. Since it will be an experimental interface, the receive and transmit lines are controlled by separate processors. The physical layer will be one of the ones used in LATM, so the interface can be used directly in our testbeds. To incorporate the interface into end systems, we will utilise planned bridges between SCI and more standard buses like VME and S-bus. The interface can then be used directly in existing end systems without a huge development effort. The ATM to SCI interface is a joint effort between Dolphin SCI Inc. and NTR. The first rough design has been finished and the first prototypes will be available in December 1993.

6 Multimedia applications

Norwegian Telecom Research' experimental application activities are formed as research programs (MultiTeam and MultiPost) on distributed multimedia applications. The intention is to outline an ambitious set of research activities in a national effort between NTR, universities and industry. The main goal was to strengthen the Norwegian Telecom's future public infrastructure to handle high speed applications, the academic research and education in information technology, industrial competitiveness and internationalisation.

6.1 Concepts to be developed

The research on distributed multimedia applications is based on integrated, collaborative multimedia desktop environments. Such an environment supports real-time audio, video, document conferencing, distributed multi-user shared workspace tools based on synchronous communication, and asynchronous multimedia document interchange. Three technology categories were looked into:

- Computer based communication, which is fully integrated with other forms of communication such as telephony, electronic mail and video conferencing.
- Shared work space area, where two or more participants can see, work and share screen and tasks. Furthermore participants can store, access, arrange and manipulate shared information.
- Group activity area, where there are support for specific tasks like chatting, brainstorming, software developing and joint editing of documents, or other customised tasks defined in demanded applications.

6.2 Testbed applications

The collaborative environment consists of multimedia desktop workstations, pen based whiteboard, high speed networks, servers, distributed applications, databases, and document architecture that will support multiple data types for sharing services.

We have identified some application categories and a number of specific multimedia applications that address the real needs of corporate users today with both asynchronous and synchronous application. First of all we want to explore new applications which may combine elements of both categories. The virtual class room and meeting room applications are examples that provide both of these elements in store and forward services and real-time video and audio to run in a distributed network. In the testbed environment we focused on the following applications.

6.2.1 Videophone and conferencing

One of the basic applications on the testbed network was implementing a Collaborative Desktop environment with multi-part videophone, videoconferencing and telephone services. New concepts of remote operation, business meetings, teamwork and remote medical treatments should be envisioned. In order to develop the proper application, we used the XMEDIA software tools and the DEC-spin conferencing software from Digital.

6.2.2 Electronic roundtable conferencing

The electronic roundtable conferencing lets brainstormers distributed on the network all talk at once on their keyboards. We have looked into the real-time meeting and conference process where documents and workspace can be shared in a co-operative task and work. The testbed was based on software tools from XMEDIA/Digital and BBN-Slate.

6.2.3 Real-time intelligent documents and task handling

Professionals need fast, accurate, team processed information handling, joint and remote editing, and tasks performed in electronic documents by accessing a number of different computers and telecommunication services. A new concept of intelligent documents have been defined where a document find its best way in the task treatments and contains all relevant media information like speech and video. The multimedia information will have different information elements like; text, drawings, spreadsheets, graphics, formularies, pictures, scanned documentation, video, interactive video, microfilm, voice and voice mail.

Collaborated studies have been done with TASKON a.s., a Norwegian software computer company in the business of work environment.

6.2.4 Electronic news and multimedia information services

New news services based on digital networks will be able to combine the depth of the newspaper with the convenience and excitement of TV. An electronic newspaper could present high-quality text next to video sequences. News articles could be received and presented continuously. Even more important, the presentation of news could be personalised according to the interests and habits of each reader. By observing the reading habits of a user, the computer will be able to create a reader's profile that can be used as a basis for selection of articles to be included in the personal newspaper. If a reader demands more information about a certain subject, e.g. by asking for more information through a hyperlink, the computer will assume this to be an area of special interest. The personal profile will be edited accordingly. The personal newspaper will contain news from a variety of sources at local, regional and international levels. Personal messages also contain important information and a newspaper will not be personal without them.

Prototypes for electronic news services have been developed using standard document definition (ODA) including text, real-time video and audio information.

6.2.5 Distance education

One of the prototype applications is the electronic classroom. It is a joint effort between the University of Oslo (USIT/UiO), the University Centre of Technology at Kjeller (UNIK) and the Norwegian Telecom Research.

In education the chalk and blackboard have for centuries been indispensable for the teacher. It is a challenge to create an electronic pen and whiteboard with the same properties as the blackboard in an auditorium. The users have a tremendous advantage in addition to write and wash out information on the board, being able to present electronic documents, audio, still pictures, high resolution graphics and video, which can be sent from any electronic class room on the network. Further more, the electronic board could be presented in one or several windows on multimedia workstations and PCs. We here envision a new educational concept where teachers can develop their class material as electronic documents which could be presented during the class and made available for the students after the lecture.

Global networks and easy-to-use computers can enrich learning within an educational environment that encourages students to work individually and co-operatively, to question facts and seek challenges. Textbooks will come in the form of computer software including text, pictures, simulations and quizzes. The material will have hyper structures to encourage exploration, and will offer access to networked databases and conferences if the student wants to explore beyond the limits of the curriculum. Electronic textbooks and encyclopaedia will have links to external databases that are constantly updated.

The electronic classroom of tomorrow will include advanced information displays controlled by computers. A large electronic board will be the natural centre of attention. Teachers can sketch on the board, show animation on it, or run a video conference through it. Students will have individual displays with pen-based input through which they can make notes, place anonymous questions onto the board, or play games. A new degree of student participation becomes possible.

The electronic classroom will be able to monitor the activities of each student. The computer will collect data on work habits, progress and performance. The data can be used to offer each student an optimised learning program based on specific needs of the individuals.

Distance education is an area that will be easier to integrate into normal education as the electronic classroom evolves. By nature, digital information is easy to distribute geographically. Linking electronic classrooms in digital networks will make it feasible for students in one location to participate in courses offered elsewhere. This will be possible without significant extra work from either side.

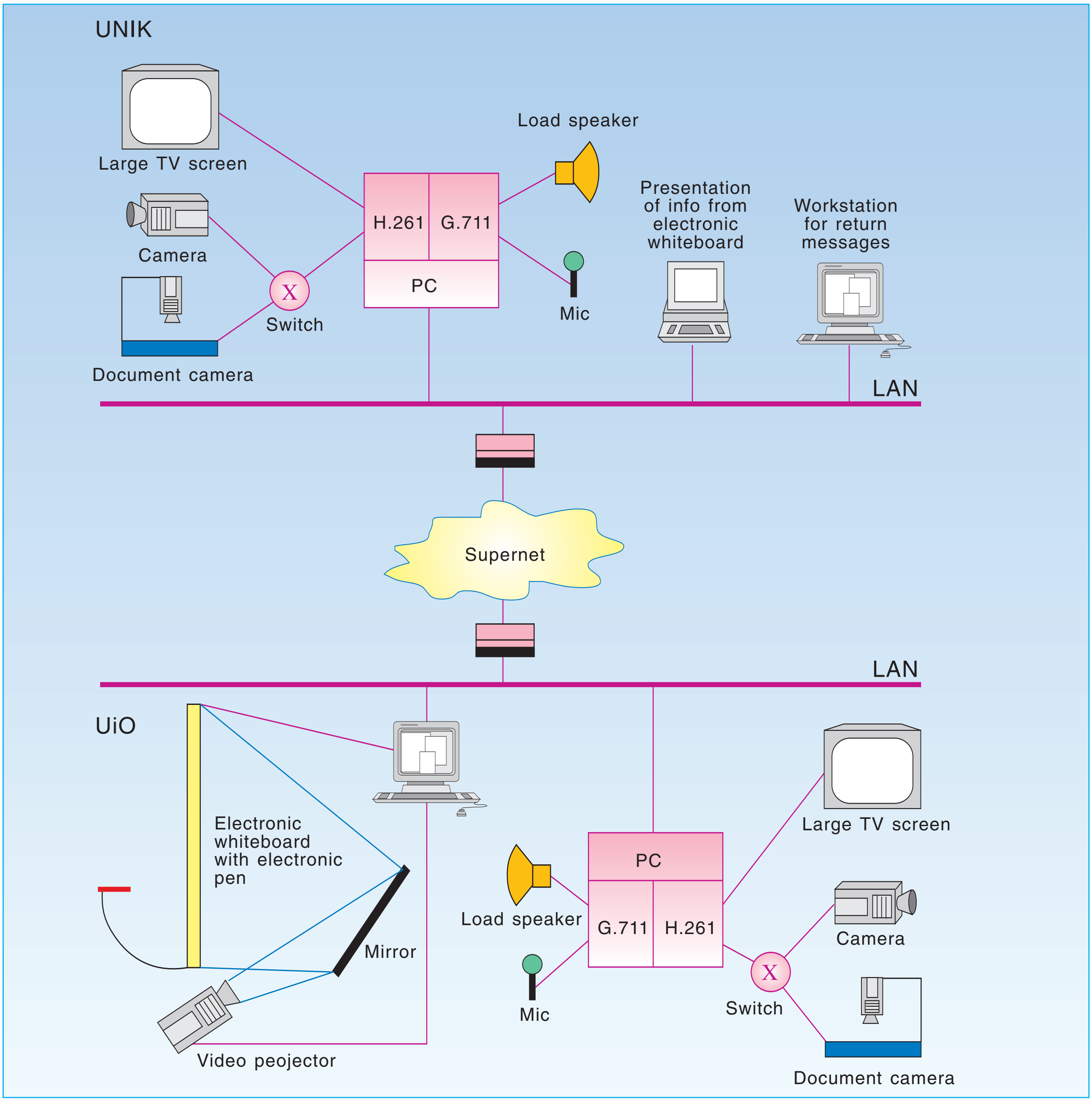


Figure 3 Set-up of distance education

The University of Oslo (USIT/UiO), the University Centre of Technology at Kjeller (UNIK) and NTR have developed and prototyped a new concept of electronic classroom. The research program was carried out in the broadband testbed lab project at NTR.

The first prototype of the electronic class room using an electronic whiteboard is now under testing for university distance education between University of Oslo and University Centre of Technology at Kjeller.

The integral part the electronic classroom is the central display and drawing area (whiteboard) which serve as medium for presentations, the set-up is shown in Figure 3. The whiteboard has a flat large semi-transparent interactive display system accessible from an electronic pen. It has a surface area of 1.5 x 2.0 meters which is divided into a pen based sensitive image, live video-window and a claim token pressure sensitive area. The whiteboard image is produced by a remote controlled digital large screen projector connected to a high performance workstation (DEC 5000) running Unix-based software. It is possible to share information with a remote location across a network. The electronic classroom could present life sized images, full duplex high quality audio, permit formal and informal interaction between remote locations and participants.

Figure 3 shows a simple overview of the main elements in the communication concept of the two collaborating universities' sites. Both sites are connected to Ethernet and FDDI networks and the 34 Mbit/s Supernet. The video signals are digitised and compressed by a H.261 codec with a compressed bit rate equal to or higher than 384 kbit/s. The pen based whiteboard requires only a low bandwidth. In our experience network throughput as low as 384 kbit/s is sufficient to support acceptable quality of the image and audio presentation.

The acoustic echoes cause acoustic instability due to the delay imposed by the network. This makes full duplex audio unusable without expensive echo cancelling equipment.

6.3 Conclusions

The electronic classroom using the whiteboard system described is the first prototype we have developed, and the work has demonstrated the potential of large screen and collaborative desktop computing as a mean for group and personal interaction over broadband networks. Future improvements of the whiteboard technology are necessary related to the optical resolution, the light output and cordless pen.

Improvements in data compression and use of high speed networks will open new possibilities to integrate audio and video with data in the network. This opens new directions and opportunities for a variety of communication applications. The most critical factor for further improvements of multimedia applications is the lack of sufficient software tools which integrate the different data types.

7 Conclusions

This article has described high speed testbeds and some of the multimedia application activities within the Norwegian Telecom. All of these are tied to either national projects or RACE projects. One of the unique aspects of the testbeds is their interconnection. In addition, the testbeds themselves are also accessible from the Internet. This offers flexibility, the ability to use realistic workloads, and access to many prototype applications. These advantages more than offset the problems created by the interconnection. The IP protocol stack is used for interconnection at the network level. It is suitable from a demonstration view point. However, it lacks functionality to take advantage of all of ATM features. In the near future it will therefore be augmented with an extension of XTP developed within RACE 2060.

8 References

1 Nicholson, A et al. High Speed Networking at Cray Research, *ACM Computer Communication Review,* vol 21, 1, 99-110.

2 Clark, D et al. An analysis of TCP Processing Overhead, *IEEE Communications Magazine,* 27 (6), 23-29.

3 Protocol Engines Inc. *XTP Protocol Definition, Revision 3.6,* 11 January 1992.

4 Fore Systems, Inc. *ASX ATM Switch Documentation, Version 2,* 1992.

5 Bryan Lyles, J, Swinehart, D C. The Emerging Gigabit Environment and the role of Local ATM, *IEEE Communications Magazine,* April 1992, vol 30, 4, 52-59.

6 The ATM forum. *ATM User-Network Interface Specification, Version 2.0,* June 1, 1992.