Telektronikk 3/4.98

Strategy in telecommunications

Contents

Feature

Special

Kaleidoscope

Telektronikk

Volume 94 No. 3/4 – 1998
ISSN 0085-7130

Editor:
Ola Espvik
Tel: (+ 47) 63 84 88 83

Status section editor:
Per Hjalmar Lehne
Tel: (+ 47) 63 84 88 26

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

Telenor

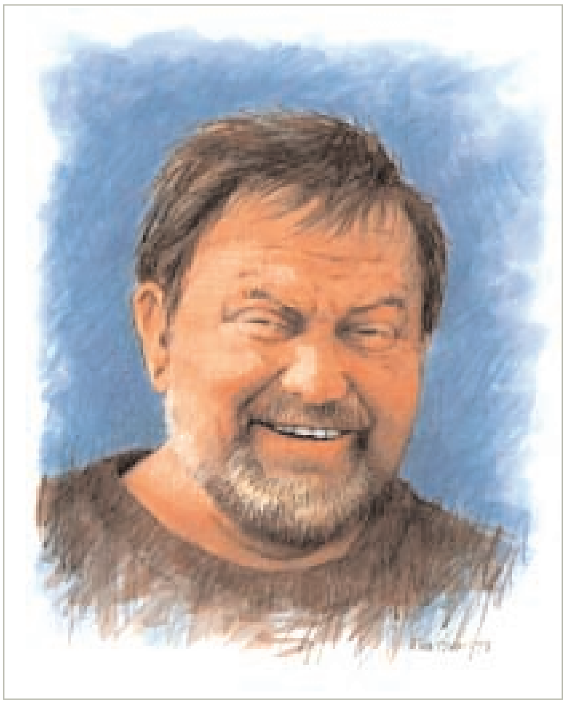
Editorial office:
Telektronikk
Telenor AS, Telenor Research & Development
P.O. Box 83
N-2007 Kjeller, Norway
Tel: (+ 47) 63 84 84 00
Fax: (+ 47) 63 81 00 76
e-mail: telektronikk@fou.telenor.no

Graphic design:
Design Consult AS

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

Editorial

JAN A AUDESTAD



Strategy has to do with the way organisations or firms are managed. Strategy as a process is not the management of the organisation itself but rather the way in which management obtains information in order to make decisions. Strategy and tactics are terms referring to the same thing and consist of methods supporting a common goal, namely the ability of a company or organisation to achieve its goals: in warfare to win the battle; in industry to seize the market. Tactics has to do with how strategic knowledge is put into operation. In this meaning, strategy is concerned with the long term planning and manoeuvrings of the company.

Encyclopaedia Britannica defines the difference between strategy and tactics as follows (in the *Macropaedia* under *Industrial Engineering and Production Management*): (1) the longer the effect of a decision and the less reversible it is, the more strategic it is; (2) the larger the proportion of a system that is affected by a decision, the more strategic it is; and (3) the more concerned a decision is with the selection of goals and objectives, as well as the means by which they are to be obtained, the more strategic it is. In this definition it is tacitly understood that the reference is a continuous scale ranging from tactical to strategic decisions. The scale is not only dependent on time – where tactical is short term and strategy is long term – but also on the quality of the decision: degree of irreversibility, part of the system affected and the number of goals concerned.

Below we are not concerned with the strategies themselves but with the methods that can be applied to identify and elucidate strategic problems. These methods are referred to as operation research. Operation research is mainly concerned with methods supporting the management of industrial organisations, governments and defence systems, and is as such a multidisciplinary science drawing on mathematical and statistical methods as well as new areas such as cybernetics, behavioural sciences, organisation theory, decision theory, game theory and nonequilibrium economy.

Operation research is a rather young science. Though operation research can be regarded as the science of applying scientific methods to management and as such has been with us for a long time, it is usually agreed that its modern evolution as a separate study began in Britain in 1937 when scientists were led to teach military leaders how to use radar in anti-aircraft warfare. The methods – or rather the selection of methods – were developed during World War II, and several operation research groups were established analysing the impact of different ways of conducting mine warfare, air raids and ground battles. The first scientific groups dealing with this new subject were formed around 1950, much because of the success of the method during the war. However, the methods were not applied at any significant scale for industrial management until the late 1950s.

In Telenor (or rather Televerket) operation research began during the 1970s. One key problem then was to use linear programming methods to find locations for exchanges and other equipment in order to make the network as cheap as possible. These are problems which are still with us.

Since the early 1990s the complexity of telecommunications has increased enormously. This has to do with several independent forms of evolution: evolution of new services in the basic portfolio of telecommunications, application of telecommunications to new areas, evolution of the technology toward processing making technologies cheaper and more versatile, liberalisation of the market opening for competition and new organisation of the business, and regulation of the competition.

As a result of the technological development society has become more dependent on telecommunications. The evolution is thus something which concerns not only the telecommunications industry but also society as a whole. The study of the impact of this evolution on society has not really begun but is likely to become a major subject taught at universities in the near future. This can also be assessed from the popularity gained recently by the conferences of the International Telecommunication Society. These conferences are not concerned with technology but with policies and interactions with society. Inspired by this way of looking at the telecommunications industry, we have started this collection of papers with a discussion of the complexity of telecommunications, where we try to evaluate several of these complexities and describe how they interact. The intention is to show that these problems cannot be understood unless proper methods are used to analyse them.

The remaining papers are concerned with different aspects of the strategic problems of the industry. These problems can only be understood in terms of the methods of operation research. The papers do not provide a complete picture of this vast subject matter but hopefully they will encourage further research of the strategic problems of telecommunications. Such studies have been done with success for the oil industry and the energy industry. The much more complex telecommunications industry has not yet been subject to this type of analysis in depth. This is an international phenomenon, and the education in business management of our own managers to be is based on studies of other and comparatively much simpler business systems.

The present papers constitute one of the first collections of operation research studies for the telecommunications industry ever published!

Telecommunications and Complexity

JAN A. AUDESTAD

1 Problems

The main problems we are facing in telecommunication are all related to complexity:

- Complexity of technology;
- Complexity of products and markets;
- Complexity of organisation;
- Complexity of co-operation or competition;
- Complexity of value creation.

This complexity is new. It did not exist a decade ago. Then the telecommunications industry consisted mainly of monopolies or cartels, the services were few and simple – and the Internet did not exist outside the academia and the research institutes. The companies offered everything to everyone within their geographic domain. When competition came, it came at a time of transition. This was a transition: from person-to-person communication to distribution of content, from vertical and tight organisation of the telecommunications companies to looser organisation and subdivision and internal competition, and from national and localised industries to multinational conglomerates forcing the telecommunications industry to form international alliances. All this has led to new and poorly understood strategic complexities.

The complexity depends on many variables and varies over time, or has spatiotemporal behaviour as it is defined in the mathematical literature [1]. Such dynamic systems may have several characteristics which make them difficult to analyse:

- They are generally non-linear;
- They will contain both fast and slow fluctuations at the same time, where some fluctuations may even be faster than the relaxation time[1] of the system;
- They may be subject to large external disturbances. Liberalisation of the telecommunication market is such a disturbance changing the basic model of the system;
- They may also be resistant to sudden changes – or be overdamped, that is, the effect of the change may come rather slowly. One example is again the liberalisation of the market, which represents a major change in one of the variables, where it may take several years before the impact is really seen on the incumbent (Sweden and the United Kingdom are examples of this). Having an expectation that damping does not exist or is insignificant, may lead to wrong regulations of the market;
- They may be described in terms of a large number of variables; that is, the spatiotemporal problem has too many dimensions for simple visualisation of it. Chapter 8, where the variables required for describing the business of providing access, is an example of this;
- They may, from time to time, show chaotic behaviour so that the future development may be impossible to predict;
- They are not in general in an equilibrium state.

Simple economic analysis cannot be used in order to describe such systems. On the other hand, advanced analytical methods such as game theory, dynamic control theory and optimisation techniques do not contain all the tools required for analysing such complex situations. This even applies to models based on computer simulation. Therefore, the problem has to be approached from different angles and at different levels of granularity, where the first objective should be to understand the underlying complexity and its dynamics. This is the main goal of this essay.

The dependability of society on telecommunications is becoming more critical. Telecommunications installations may become the target for terrorism, not only by physical attack but also using logical bombs and other ‘soft’ means. This is an aspect which has not been fully understood by politicians and regulators. Competition in a market with shrinking margins may cause reliability to be bargained against lower prices. This is one lesson learned from the Internet: much of the success is related to low prices and high functionality. However, reliability and quality of service is low.

Another experience is that the quality of service in terms of reliability, response time and stability is poor also for local data networks. The reasons for this are two-fold. First, quality of service is bargained against low price as for the Internet, and the customers – residential or business – are sensitive to price but have little awareness concerning reliability. Second, the way in which administrative data systems are specified and implemented is different from that of telecommunications equipment: in the latter, most of the effort is put in handling of exception conditions, autonomy, automatic fault recovery and fault isolation. Telecommunications systems have also strong real-time and concurrency requirements. The real-time requirement implies that all processing must be done within strict time limits. These considerations are almost absent in administrative systems. Solution of real-time problems require use of dedicated modelling methods for telecommunications software reflecting the dynamic behaviour of the systems. Concurrency means that thousands of simultaneous usages of the telecommunications software may exist in the same exchange. Because of the real-time requirements this implies that each usage requires a separate copy of each software module and that the execution of these copies are not synchronised. These considerations make the design of telecommunication software a separate – and complex – discipline of computer science.

This may have an impact on the convergence of telecommunications technologies and information technologies making this convergence not so trivial. Moreover, the data industry is entering the arena of the telecommunications industry producing equipment for switching of traffic. Since these two industries have different production philosophies, this may change fundamental requirements of telecommunications on reliability making the network cheaper but less dependable. This aspect is probably one of the most important elements of a *technological* strategy for telecommunications.

The markets are also changing. Most of the services enhancing the basic telephone service are new. They became universally available less than ten years ago. The age of services requiring the intelligent network technology is less than five years. These developments are the basis for the distribution of content rich services. Since we have had these services for such a short time, it is very difficult to estimate how they will evolve in the future: they have no history to base forecasts on, the technology on which they

[1] *The relaxation time is a measure for the time required for a system being brought out of equilibrium to again settle in the same (or a new) equilibrium state if it is left alone.*

are built are evolving and changing rapidly, and telecommunications has become a competitor to other distribution channels of many products (newspaper, trade and so on) and it is difficult to assess the persistence of traditional channels: a change will come for certain, but the problem is when.

The liberalisation of the telecommunications market is changing the way in which the business is organised and conducted. The monopolies were vertically organised companies offering all aspects of telecommunications to everyone. Competition is now met in small niches of the business as well as from vertically integrated companies mimicking the incumbent. This is making the industry complex and requires a new look at how value is really created within the industry. In the core business the markets have been big with good margins. This is changing since competition has forced price reductions on core products. Many of the new markets have not materialised yet. This leaves the industry facing a shrinking primary market and a new market with much potential but high investment risks.

The new situation requires that the telecommunications operators are reconsidering how value is created in the organisation and when it is advantageous to co-operate or compete with newcomers in the marketplace. These are difficult problems and may require development of new strategic methods.

These problems are considered below, where the main concern will be to assess the complexity all these items will impose on the industry. Access operation will be used as an example for illustrating how big and complex a strategic analysis is, even when it is confined only to identifying the parameters of one particular part of the industry.

2 Complexity and Technology: Does new Technology Reduce the Time to Market?

The evolution of telecommunications has been rather slow but is accelerating. By slow we mean that the time between major events and the time to invent new technologies have been long, in general more than ten years. One important issue is then to answer the question: has this time been reduced significantly during the last few years? The surprising answer is that this is not so in general, though there are systems and methods which do just that in particular circumstances. This is the core of the problem we are facing today: there are two paths of development of telecommunications where one is slow and the other is fast. The fast developments are associated with new products on the World Wide Web and new services on intelligent network nodes; that is, development of products which can be made from modules and tools existing on the platform on which it is implemented, such as html on the Internet and service independent building blocks (SIBs) on the IN. The development of the platform itself takes time, and is an example of the slow path of development.

As an example, it took more than ten years to develop the IN technology; in contrast, it takes only days to develop a service like freephone on this platform. This duality is becoming a strategic problem in telecommunications: it has changed our expectation concerning the time it takes to bring a new product to the market, and it causes us to underestimate the cost of radically new products as compared to products which can be designed on ‘smart’ platforms. We also often overlook the fact that there is a continuous evolution of these ‘smart’ platforms, which means that the two development paths often coexist. This means that we also often have the choice of either implementing a service or application on the existing platform, or wait until a new version of it is in place. In the first case we may develop a service which is not optimal for the market; it may even be detrimental to it. In the second case the product may come too late to the market because other developers have come up with replacement products, or the market has changed the focus towards other products.

Take the universal personal telecommunications (UPT) service as an example of this problem of duality. The service is marketed in Norway as ‘Alfanummer’. The idea behind the service is indeed a smart one: the service offers mobility between any termination – fixed or mobile – in the network on one number and one service profile of the user. The service is easily – and cheaply – implemented on the IN platform. However, since this platform does not offer a man-machine interface to the user capable of offering protection against fraud, simple handling of service menus, and memory of long dialling sequences, the service is difficult to use. What is required is not a better IN platform but a new platform replacing the telephone apparatus at every home and industry by new equipment offering simple interfaces based on smart cards, displays and automatic handling of complex menus. On the other hand, GSM offers just this type of interface, and because of its widespread use it is replacing the need for UPT.

Let us now review some of the major events leading up to telecommunications as we perceive it today.

The first major event in telecommunications was that Strowger invented automatic switching in 1889, leading to the first commercial electro-mechanical switch in 1898. Although the technology offers cost reduction for the operators, it took 75 years to automate Norway completely. And Norway is no exception in this evolution.

Almost forty years after automation of telephone switching the first automatic systems for transfer of text were invented. In 1933 we got the first telex network in Germany, though this service did not become generally available until after World War II. During the 1960s the first communication satellites were put into orbit, experiments with data transmission began, and the first data controlled exchanges were implemented. The background invention enabling this radical evolution was the transistor. These systems required high reliability and small size which could not be achieved with the vacuum tube.

About 1980 we got the first automatic mobile systems. These included NMT, AMPS and TACS[2] for land mobile applications, and INMARSAT[3] for communi-

[2] *NMT (Nordic Mobile Telephone) is the Nordic automatic land mobile system operational from 1980. AMPS (Automatic Mobile Phone System) is an American land mobile system put into operation during the first half of the 1980s. TACS (Total Access Control System) is a British adaptation of AMPS put into operation in 1983.*

[3] *International Maritime Satellite Organization inherited the MARISAT system from Comsat General and commenced operation in 1980.*

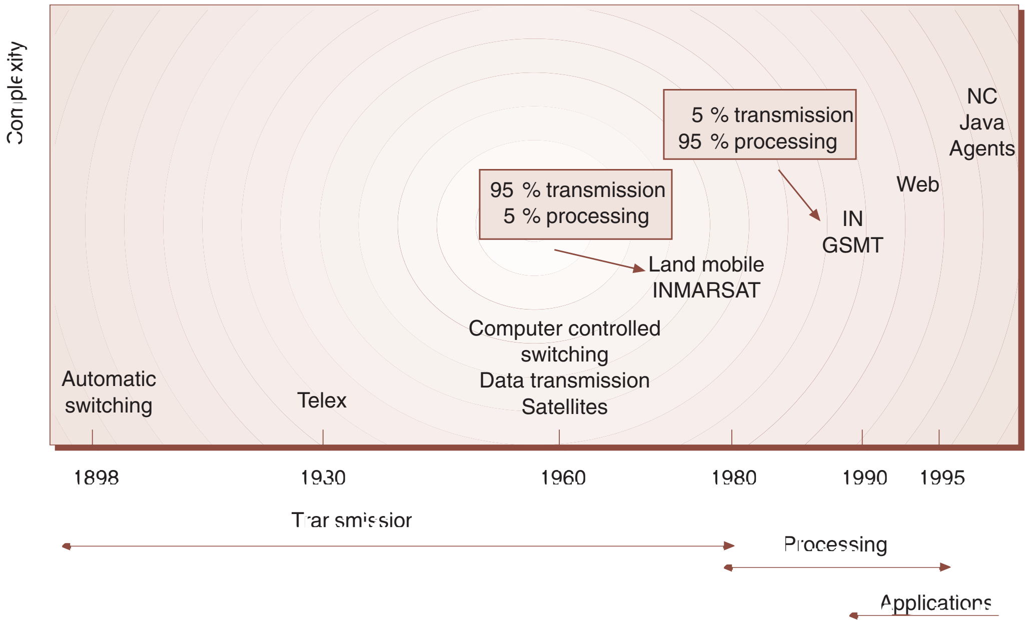


Figure 1 Evolution of telecommunications

cations with ships. None of these technologies would have been possible without the microprocessor. After 1990 we have seen an explosion of new technologies: intelligent networks, GSM, World Wide Web, network computer, Java, intelligent agents, value added services, and value added networks. It may look as if the complexity of new systems invented at any time is proportional to the complexity of the previous systems taken together, giving rise to an evolution which is exponential – often referred to as natural evolution. This evolution is illustrated in Figure 1.

Until about 1980 the main focus in telecommunications was on transmission. After 1980 the focus has been on processing. By processing we mean activities related to implementation of both services, applications, protocols and platforms, and complex procedures like cryptography, channel coding, voice coding and error protection coding. In advanced products like GSM, demodulation of radio signals, voice coding and channel coding are specified and implemented as processing activities. In 1986 the first simulations of channel coding and demodulation of GSM were performed on a Cray computer. This was the only machine at that time fast enough to do this using software only. Five years later the processes were easily implemented on microchips.

In Figure 1 the change of focus is illustrated by how much of the specifications of the INMARSAT system and of the GSM system were concerned with transmission and processing. During the decade the focus changed completely. The change of focus has also resulted in a tremendous increase in complexity during this short period of time. The increase in complexity of the systems is owing to the increased capabilities of the microprocessor: procedures and algorithms that had to be implemented in hardware are now implemented in software, and procedures that could not be implemented at all or were highly impractical are now possible as computer solutions because of the low cost of microchip computers, advanced development tools for software production, and more processing and storing capacity per unit of area. One example is the echo canceller used to suppress echoes returned from the terminations of the four-wire telephone network. In 1970 these devices had to be about one cubic metre big in order to accommodate the lengths of coaxial cables and the number of magnetic switches required for constructing the transversal filters of the canceller. Ten years later this bulk was reduced to a small printed circuit card containing a few microcircuits and a few other components. Ten years later still the whole device was contained on one microchip not distinguishable from any other microprocessor. The technology of echo cancellation has during these twenty years developed from pure hardware construction to pure software construction. This example is typical for much of the evolution which has taken place during the last twenty years. In addition, the use of software technology has increased the number of applications which can be constructed.

The focus has again shifted during the 1990s. Because of the web, the designer may now turn the main focus towards the applications and less towards the technology enabling them.

However, it should be noted that developments that involve new software methods or hardware still take time – from five to ten years or more – so that the development time of such solutions have not been reduced significantly. This applies, for example, to the development of nodes offering intelligent network (IN) capabilities and enhanced user capabilities (CTI[4]), software platforms like the World Wide Web and CORBA[5], and new protocols such as IP6[6]. However, applications built on platforms already implemented such as intelligent network nodes, the World Wide Web and CORBA can often be developed in a matter of days or, at most, during a few weeks. This has created an imbalance in the development time of new products:

Platform products	*Development time*
Intelligent network (IN) node	10 years
GSM	10 years
World Wide Web	5 years
CORBA	10 years
ISDN	15 years
Computer telephone integration (CTI)	10 years
Derived products	
Freephone	1 month
Universal personal telecommunications (UPT)	6 months
Premium rate service	1 month
Web page	1 week
Electronic newspaper	2 months
Transaction services	2 months
Products with enhanced technology	
Centralised queue	1 - 2 years
Intelligent building	1 - 2 years
Wearables: experiments	months
Werables: product adapted to market	several years

Figure 2 Development time of different technologies

those which can be developed on such platforms and those which cannot. The first type of product will have a short development time and, hence, a short time to market, while the second type of product will have a much longer time to market. In many cases this leads to unrealistic expectations concerning production time and resources required to implement a given solution. This applies in particular to products where most of the product can be made on such a platform but some part of it needs to be developed. One example is the implementation of the centralised queue service which can be implemented on an IN node in a very short time but requires new development involving remote procedure calls between the PABXs and the IN node. For the intelligent building the situation is similar: the product can be implemented quickly on existing platforms.

The price tag will be high for such solutions, while it takes time to develop the cheap technology.

These elements are summarised in Figure 2.

Technology is thus an example of an evolutionary variable in the dynamics of telecommunications containing two (or more) types of temporal cycles: short cycles and long cycles. This may also give rise to different behaviour of competing companies depending on where they are in these cycles. One example is that companies focusing mainly on the Internet may be in a period with short cycles concerning the development time for new products. Other companies may be in a period of predominantly long cycles building up the technology which may make them change to short cycles in order to compete with the Internet. This is illustrated in Figure 3. The profits may as a result also be cyclic and may correspond to a special kind of leapfrogging [2] which may be common in telecommunications in the future. The competition between such companies is not equal – and cannot be made so by any regulatory interference.

If we look at the technological evolution we may divide it into the following main phases [3], [4]. Before 1980 the focus was chiefly on transmission. This was the great era of multiplexing technologies and source coding techniques for both speech and picture. The 1980s were the years of processing where the microprocessor led to the PC revolution and enabled the design of equipment for mobile

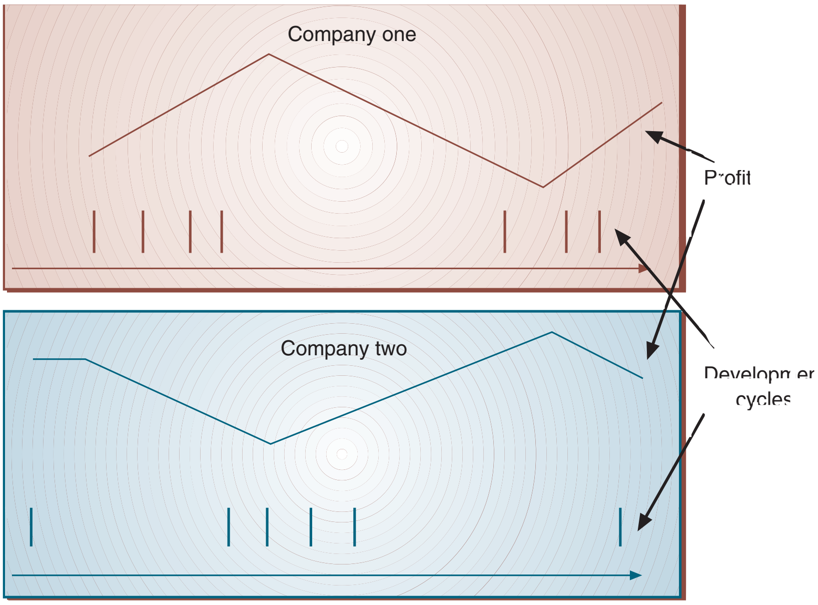


Figure 3 Product cycles

[4] *CTI is equipment combining telecommunications and computer applications such as deriving customer information from the called number. It is an abbreviation for Computer Telephone Integration.*

[5] *CORBA is an abbreviation for Common Object Request Broker Architecture, and is a new industrial standard for platforms for distributed processing.*

[6] *Internet Protocol version 6.*

communications and control devices in automobiles. The PC did not lead to more traffic in the telecommunications networks. The PCs wre isolated machines in the homes or connected to closed networks in companies. The 1990s have been the era of access to information manifested by the World Wide Web and the search engines. The enormous increase in traffic caused by the Web came about because the PCs were there as a result of the previous development. According to [3] the next decade will be the era of interaction where micro-miniaturised sensors and remotely controlled devices will be the basic components. One example is the video camera now reduced to a microchip with the micro-miniature lens glued to or integrated within the chip. The price is about $ 10, and will be reduced significantly during the next few years. Similarly this development may enable micro-miniaturised radars, chromatographs, pumps, valves, motors and so on, revolutionising many current technologies such as medicine, environmental control, process management and surveillance [5]. This technology will indeed require telecommunications! But telecommunications will most likely change in order to meet new requirements with regard to bandwidth, ubiquity, pricing, and so on.

Of all these phases, the telecommunications industry has only been in the driver's seat for one of them: the development of transmission systems. The other areas have come as a result of developments in the data industry, the information industry, the microelectronics industry and the academia. This is also what makes the telecommunication industry so vulnerable: its field may easily be invaded by companies skilled in these new areas. Still we are struggling with the convergence of telecommunications and information technology while the new development may require other types of convergence completely unknown to us.

3 Complexity and Dependability: The Impact on Society

The number of main switching nodes in the telecommunications network has been reduced considerably during the last ten years, and there is still much potential for further reduction. In Norway there is now 180 main switching nodes in the telephone/ISDN network. One of the GSM networks in Norway has about 40 % as many users as in the fixed network. These are, however, served by only eight exchanges. What has happened is that each node has become larger, and a failure of any such node will affect larger parts of society and require an increase in reliability of each node of the order of one or two in magnitude in order to keep the risk level at a constant value. In addition comes the general increase in capacity of optical fibre systems (order of terabits/s) creating a network with overwhelming overcapacity where there is little concern in bringing calls over long distances and thus wasting capacity. This means that we can reduce the number of switching nodes by connecting users over long distances to the few nodes which are left.

In this way the network may be made smaller with regard to the amount of equipment required. However, this may not result in a proportional reduction in cost since the reliability must be increased in order to keep the effects on society small. One good example is the cost of satellites. They are a comparably simple equipment, but the cost is tremendous because of the reliability requirements which must take into account the harsh environment and their inaccessibility to repair.

The sensor technology of the next decade may require new usage of telecommunications in other areas of society increasing the dependability between telecommunications and vital functions of industry, governments and individuals.

The dependability of the society on telecommunications is increasing and may already be critical in certain areas, amongst others the impact on our economic systems. One unanswered question is how long the New York Stock Exchange and the Tokyo Stock Exchange can be without telecommunications before the world enters into a severe economic crisis. Another question is how quickly the air transport systems of the world will break down as a result of failure of a few critical nodes. Will telecommunications become the new targets for terrorists in order to hurt the society? Will they use physical attacks or timed logical bombs? Will such bombs be implanted at the time of construction or inserted later via insecure accesses to the network's operation systems?

A measure for dependability is suggested in [6]:

$$D = \alpha \log(NT),$$

where the measure for dependability is D, α is some constant of proportionality, N is the number of customer circuits affected and T is the total downtime in hours. If we put $\alpha = 1$, this measure is in many ways equivalent to the Richter scale for earthquakes (however, do not bring the analogy too far):

- If the value is less than 6.0 the impact may be severe but not critical. This will happen if one million circuits are affected for one hour, or 100,000 circuits are affected for ten hours;
- If the value is above 7.0, the impact may, just as for earthquakes, be disastrous. [6] estimates that the outages in the USA in the summer of 1991 were of this order.

Using a value of $\alpha = 1$ may apply for very large networks such as the whole international network. A larger value of α may be used to estimate the impact on smaller networks. However, we should not put too much significance into this formula: it is merely included here to visualise the general problem of dependability.

Values above 7.0 are rare but their probability may be increasing for several reasons:

- The increased complexity of telecommunications networks;
- The tendency to centralise nodes as explained above;
- The increased need for frequent upgrades of software and hardware of nodes in order to meet market demands for new services and reduce operation costs;
- The reduced capital available for investments and maintenance because of the liberalisation of the market resulting in competition between operators bargaining reliability for price.

Another problem is related to the number of error reports created by the network. In a network with n nodes where all the nodes are connected with one another the number of error events generated if one node fails is also n. With a mean time between failures of a node of t in years the mean time between failure of the system will be t / n, and the number of error reports created per year will be n^2 / t.

With a mean time between failures of 1 year, the number of reports created in a network consisting of 1,000 nodes will be one million per year, or about 2,700 per day. In a large network consisting of 100,000 nodes, the number of reports per day will be about 27 million. Even if the connectivity of the network is only 1 %, the number of reports per day will be 270,000 in such a big network. The latter number should indeed be taken seriously when considering the reliability of the Internet which already consists of a large number (several millions) of nodes (servers) with rather low reliability.

Hence, we see that the reliability of the network may become a cost driver since every error occurring in the network requires maintenance activities. In order to reduce the number of points in the network requiring maintenance, the number of nodes should be made smaller, that is, increasing the size of each node. In accordance with the above consideration this may increase the dependability of society on the network so that the two strategies may be in conflict with each other.

4 Complexity and Services: New Ways of Making Products

Until the beginning of the 1990s, services comprised mainly simple two-party calls. The main focus was the residential market. Since then we have got supplementary services such as call forwarding, call barring and call waiting and virtual private networks, and, most important, the Web on the Internet. The Web provides us with a completely new concept of what services are. The network computer, the intelligent agent and the Java language are derivatives of this technology. Note that the development of advanced telecommunications services and the derivative technologies of the Internet took place at the same time. This has created two strategic problems in telecommunications. First, it seems as if developing solutions using Internet technologies is much faster than using telecommunication platforms, and second, that the products derived from the Internet platform are cheaper and better than the equivalent products derived from telecommunications technologies. This may be true in special cases but is not true in general. As was shown above, the difference in the time to manufacture a new product depends not on whether it is done on a telecommunications platform or on an Internet platform. The difference lies in whether all components r-equired for the product are available at the platform or need to be developed first. In the latter case, the development time and production cost are more or less the same for the two types of platform. There may, however, be a qualitative difference between the products. We will come back to this in Chapter 7.

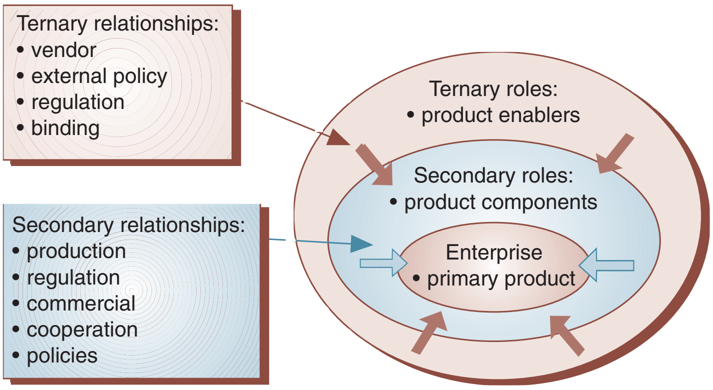


Figure 4 Product hierarchy

Services are just as much bundling and combination of offers as single products. One example is bundling of TV channels, local advertisements and subtitles over a satellite network involving several independent business systems. Electronic trade combined with payment services, security and anonymisation services, physical delivery of goods, delivery of telephony, delivery of electronic newspapers and advertisements is another example. Electricity, cable TV and telephony may also be combined as a single offer and on one bill. Services may be delivered by one company or be composed of services delivered by several specialised companies. Alliances between different actors may be built on single service offers or on portfolios of services and other products, resulting in interesting and difficult games of importance, dominance and buy-outs. This type of complexity is treated in [7].

Telecommunications services for the business market may be used by the companies purchasing them as ‘raw’ material or as components for building their own products. In this way telecommunications may become a secondary or even ternary product in this production. Electronic

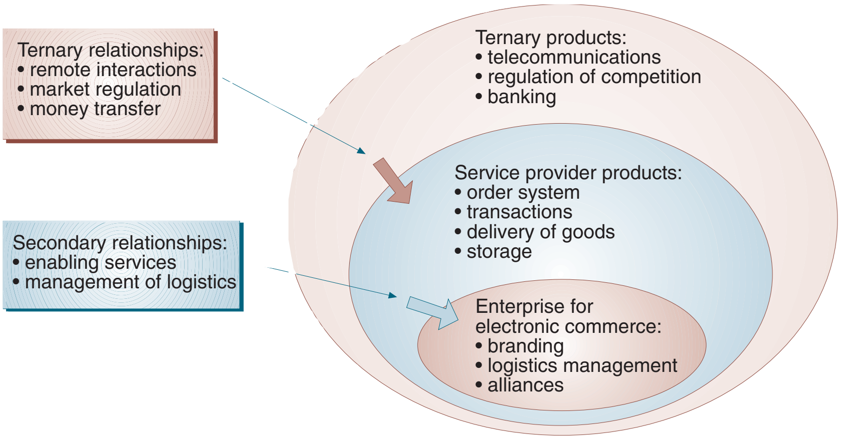


Figure 5 Product hierarchy: electronic commerce

banking, electronic newspaper and electronic commerce are examples where telecommunications services are just enablers for the primary service. The telecommunications product may enter the value producing system of the company as an element of or as a support activity in a value chain, as a layer in a value network or as a skill in a value shop [8]. The product is different in each of these cases because it will meet different interfaces. The impact of this is poorly understood by the telecommunication operators and is probably critical to them. In this market it is difficult to analyse what the product really is and how it is viewed by the buyer, and how it will evolve.

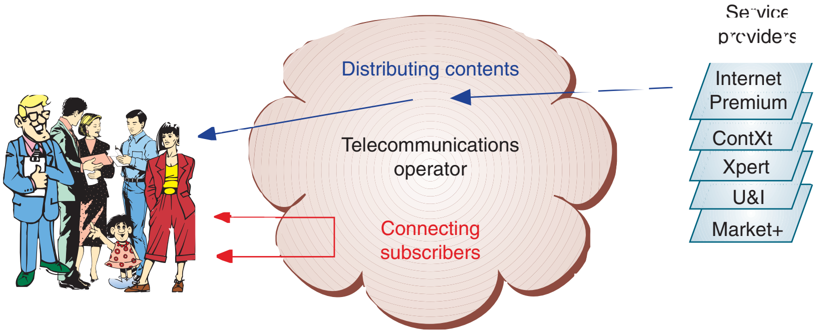


Figure 6 Interaction with service providers

Figure 4 is an enterprise model where it is shown how a three stage hierarchy of 'influences' may be incorporated in the production chain of the enterprise under consideration (the inner oval in the figure). The principles are equivalent to those defined for distributed systems [9] but slightly modified in order to be applicable to systems which are more general than distributed computer systems (see also [10]).

The enterprise produces a primary equipment for the market. This product may require resources and elements obtained from secondary sources. The product may also be directly influenced by regulation and other influences external to the enterprise. The ternary relationships operate on the secondary elements of the product and may not be visible to the enterprise. Still they may influence the product.

Figure 5 illustrates these principles for the electronic commerce. The product components provided by the primary enterprise may be branding, logistics and building of alliances. Service providers deliver secondary products such as order systems, transaction support and delivery of goods. The business of the service providers may then be influenced by the regulation of the competition, make use of the banking system, and use telecommunications as an enabler both for the remote ordering system and the transaction support system. For enterprises living from the electronic commerce it is not important how this is done. For the enterprise it is more important that payment transactions are managed. It is not important to know how and by whom this is done. This is exactly the role telecommunications has with regard to electronic payment in shops (point-of-sale systems). In this example telecommunications is a ternary product. If the goods were delivered electronically (music, video, pictures), telecommunications may have been upgraded to a secondary product in the delivery chain. In this and similar examples it will never be upgraded to become a primary product.

5 Complexity and Market: New Ways of Doing Business

Another way of looking at telecommunications products is shown in Figure 6. There are two markets: one market where telecommunications is the primary product, and one market in which the telecommunications operator distributes contents from service providers to people and machines connected to the network operated by the telecommunications operator. The traditional business of telecommunications is to interconnect subscribers, and the main product has been to offer two-party calls between people and between machines. In the new market the business idea is to use telecommunications for distribution of content. The telecommunications products may then enter the production chain of the customer as secondary or ternary products as explained above. The telecommunications operator will see the content as emanating from the service provider (in the figure indicated by fictitious names Internet Premium, Xpert and so on). For the service provider telecommunications is just one way of distributing the content in order to reduce cost and reach new customers, and it is only secondary to the main product which is the content. The main value the telecommunications operator can offer in this configuration is the number of subscribers connected to the network who

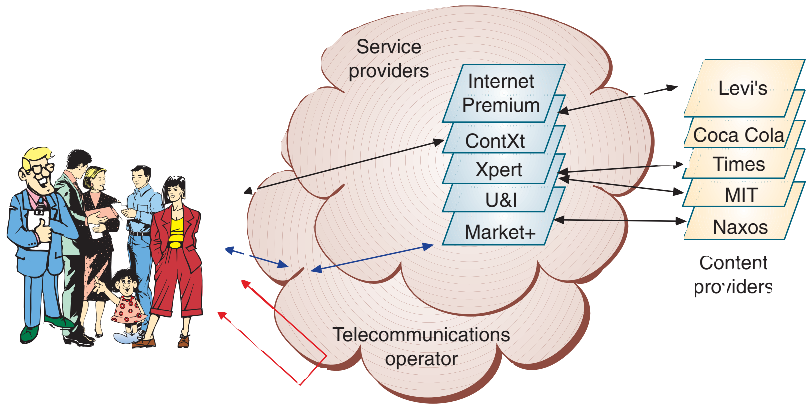


Figure 7 Interaction with content provider

are potential consumers of the content. In Figure 7 this picture is taken further because the source of the content may not be the service provider but some other company such as Levi's or Coca Cola providing electronic commerce and advertisement, MIT offering distance learning, Naxos offering music, or The Times offering an electronic newspaper. The service provider then offers storing of information and processing of support activities such as payment, security, electronic marketplace, integration of contents and packaging. Telecommunications as a part of these products is hardly seen by many of these content providers except as a cost factor, and possibly indirectly in terms of the efficiency of the service provider.

This new market is not mature; it is developing now, and nobody knows how it will develop and when it is possible to earn money from it (see for example [11] and [12]). The problem is associated with the duality of development time described above (see Figure 2). Most of the required functionality is in place on the Web or other commercial platforms for 'fast lane' development. However, we still lack some functionality such as connecting these platforms with cash registers, databases, sensors and local processors, and to integrate them with systems for logistics, transaction support and security. This requires development in the 'slow lane' in order to bring forward these new basic technologies.

The markets for traditional telecommunications are probably stagnating soon. For telecommunications operators it has been very difficult to assess whether the market is increasing, is stagnated, or is decreasing. The reason is that it is difficult to distinguish between different types of traffic in the traffic machine: telephony, mobile communications and Internet. For telephony it is also difficult to distinguish between voice traffic, telefax traffic and other data traffic. There are two other factors adding to this difficulty:

- Traffic is moving from traditional telephony to the Internet because the latter offer better functionality. The best example of this is telefax traffic which is replaced by e-mail. E-mail also replaces other telephone traffic because it is more convenient to write messages rather than speaking to an answering machine;
- The other reason is cannibalisation; that is, one service is replaced by another. The best example is mobile telephony which for a long time has been supplementary to telephony, and has been a major reason for the general increase in telephone traffic. However, with the penetration of mobile telephony we now have in Norway, it is starting to cannibalise the market for fixed network telephony.

The point illustrated here is also related to C4 convergence. The four Cs stand for content, computer, communications and consumer electronics representing each of the four industries required in order to bring information to the user over electronic media. They may form an hierarchy or production 'chain' as shown in the left part of Figure 8. Each industry will also perform given tasks as shown in the right part of the figure.

The chain has consisted of independent industries delivering their specialised products. As content is becoming more and more important, the interaction between the four industries is increasing. They are also becoming more dependent on one another with regard to type of product, interactivity between consumer and provider, quality of delivery and size of consumer market. This has led to new types of alliances, buy-outs and co-operation along the chain in Figure 8. This is a game of gaining advantages:

- The content provider will ensure access to good communication channels and avoid competitors gaining access to the same channels and the customers connected to them;
- The manufacturer of consumer electronics may join with a content provider and make equipment dedicated to displaying the content for that provider in order to increase his own market;
- Computer manufacturers may provide dedicated equipment for processing certain types of content and join with a communication provider in order to obtain a sizeable market for that product;
- Communication providers may buy content providers in order to get exclusive rights to distribute the content;
- and so on.

Several attempts on implementing C4 convergence in this way have been made, but in general they have not been successful. Either this type of business integration is too difficult or it turns out that there is really little economic advantage in doing so. Another aspect is that electronic media are not mature and are changing so rapidly that alliances which are good one day may be bad the next day. C4 convergence may become important when this part of the business is stable. A survey of the opportunities and the expected size of the markets are found in [13].

6 Complexity and Organisation: Co-operate or Compete?

Above, we looked at the complete chain of activities from creation of content to displaying it to the user. In this chapter we will look more closely at the communication part of the C4 chain, showing that it alone consists of a large number of actors or activities.

We have been used to regarding telecommunications as one homogeneous and integrated business. This picture was in many ways true until the monopolies in telecommunications were resolved: the monopolies provided, as one single company, all types of telecommunications to everyone. In data transmission and mobile communications the disintegration of the system into several roles has been emerging during the last twenty years. However, these businesses have previously been regarded as being too small to be representative of the evolution of the business. Electronic distribution of content and the explosive growth of the Internet are other tendencies which have caused the emergence of new roles

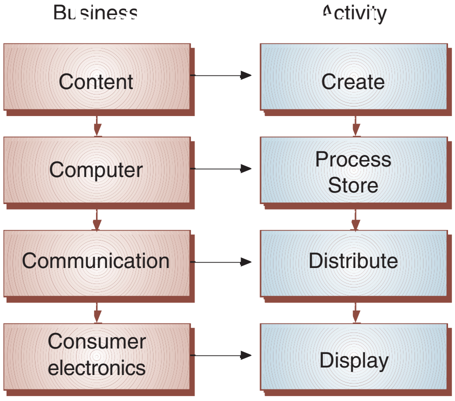


Figure 8 C4 convergence

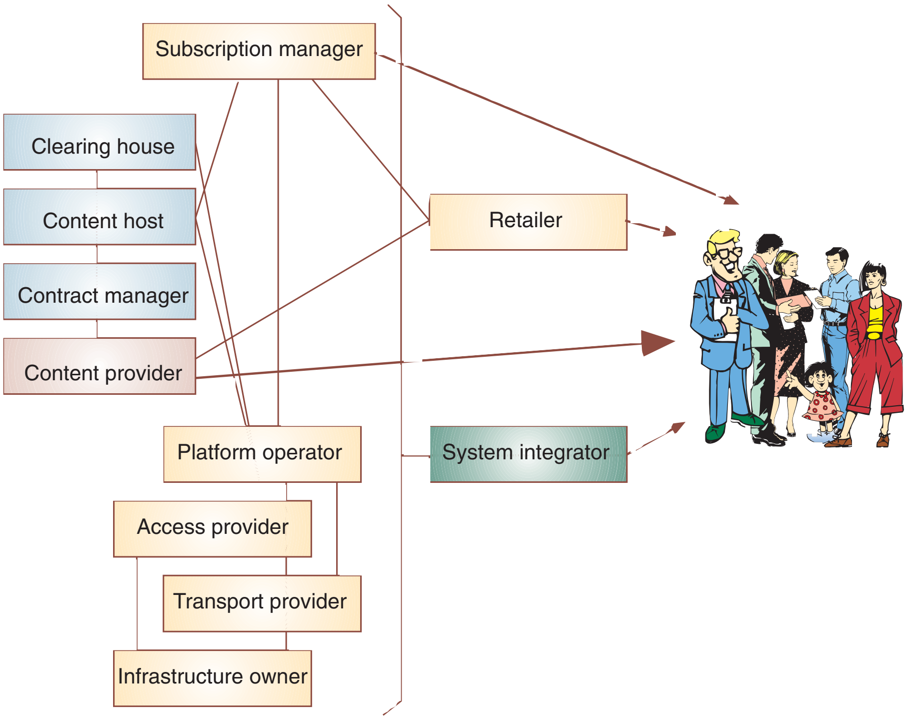


Figure 9 Business system

and new businesses in telecommunications. Still these new businesses are small as compared to the traditional businesses of telecommunications companies.

In a model developed at Telenor the telecommunications business is divided into ten roles plus the role of content provider. The content provider role, as illustrated in Figure 7, is in itself very complex and consists of a large number of different and incomparable businesses. We have included the role in our model for completeness. The different roles are interconnected with one another giving a model as shown in Figure 9. The model in the figure is not complete for two reasons: it shows only some of the possible interactions between roles, and it contains only one instance of each role. Several roles of the same type may be present in the same company, for example serving different market segments or services (one instance of the role serving the residential market and one instance serving the business market, or one instance serving the market for fixed network services and another instance serving the market for mobile services).

Another important aspect determining the composition of the model is based on the observation that for each role there are in fact companies doing separate businesses there. Note also that the roles are atomic in the sense that it is difficult to divide them further without getting too specific and losing the genericness of the model.

The TINA Consortium[7] has developed an equivalent model [15] for telecommunications. Their choice of roles is purely technical and its main focus is to determine which technical interfaces need to be defined. Our model is more focused on the commercial issues and therefore this model contains a richer structure than the model of the TINA Consortium.

The roles are as follows:

- *Infrastructure owners* own buildings, ducts, pipes, masts and other basic infrastructure required for accommodating cables and equipment. Most existing telecommunications operators are in this role but it also contains other parties such as road authorities, gas works, railways, cities and electricity companies. One important aspect of many infrastructure owners is their right-of-way which enables them to put up supporting infrastructure efficiently.
- *Transport providers* own cables, fibres, frequency licenses, satellites and so on, and offer bandwidth, quality of service, capacity, multiplexing of signals and synchronisation of networks.
- *Access providers* offer interconnection of users of telecommunications and networks. They own technologies like copper access lines, fibres, concentrators, frequency licenses, base stations, satellites, satellite earth stations, cable TV networks and so on. In Chapter 8 we will look in more detail at the immense complexity associated with the access network and how different techniques are changing from being supplementary to becoming replaceable. This will serve as a good example of the increasing complexity of the system.
- *Platform operators* own switches, servers, processors and 'intelligent' nodes for supporting processing, storing, switching, interconnection and routing of calls and content. Note that a platform may be any type of processing device. It may be localised or distributed. What is to be understood by a platform is not how it is designed or what it is called, but what it does. The platform may be public or private. The platform operators are using equipment obtained from the transport providers for interconnecting the machines making up the platform and technologies of the access providers for connecting the users to their platforms.
- *Subscription managers* keep customer relationships, are responsible for billing, offer customer care, and interface the market. The customers may be residential customers or businesses.
- *Retailers* sell subscriptions, traffic and customer care to the market and may

[7] *Telecommunications Information Networking Architecture Consortium (established 1993) is a co-operation between telecommunications operators, telecommunications manufacturers and computer and software manufacturers with the aim of specifying the next generation distributed platform for telecommunications applications [14], [23].*

be a market interface representing one or several subscription managers.

- *Content hosts* offer storage and processing to content providers.
- *Clearinghouses* offer transaction support, security, information filtering, trust and insurance against fraud and liabilities.
- *Content providers* create and bundle information and may be any type of company or business offering content such as newspapers, universities, merchants, public offices and physicians.
- *Contract managers* are in many ways equivalent to the subscription managers. However, their main asset is to manage the contract between the content providers and the providers offering services to them. This will involve payment and billing services.
- *System integrators* design solutions for customers and bundle technologies and contents from other actors of the model. These are roles which can be filled by consultant companies, computer industries and even advanced content providers. The system integrators do not need deep technological skills or component expertise. They need the skills of the entrepreneur and should possess competencies like consultative sales, knowledge of the value production system of the customer, overview of available technologies which may be used in the solution and their relative merits, insight in the expected evolution of the technology and the market, and ability to develop new solutions.

The first six roles are those of traditional telecommunications. The other roles are new and have existed for about five years or less, and are still in their embryonic form.

Note that manufacturers of equipment are not included in the model. The reason is that the model contains only roles which may take direct part in producing the telecommunications product and have it delivered to the market. In this respect, the equipment manufacturer is not visible and plays no role in the production of services. Of course, a manufacturer of equipment of software may participate in producing telecommunications services but, if so, in one of the roles defined above. Examples are IBM as system integrator and Microsoft as platform and access operator (if they realise the Teledisc system).

There will be four generic interactions with the market. These are emanating from the subscription manager role, the retailer role, the content provider role and the system integrator role. These interactions may also be simultaneous and have different content in different situations. The content of these interactions may be:

- The subscription manager is responsible for marketing the overall product, doing customer care and billing, and being the point of contact between the customers and the rather complex structure producing the services;
- The retailer sells terminal products (television sets, PCs, telephones etc.) and manages the sales of subscription contracts;
- The content provider may have any type of interaction with the market. This issue is too complex to consider here, and the content of this interaction is not important for understanding telecommunications. As will be discussed below, the important point is that it exists at the same time as the other interactions and may be independent of them;
- The system integrator may have contacts with the market when co-ordinating the establishment of a system or connecting a new customer to it. This situation is common in the business market where the initial sale of a solution is undertaken by a system integrator. Another more continuous contact with the market may be to own the brand name for certain products.

One strategic challenge in telecommunications is to identify which type of organisation each role represents and how value creation takes place in the organisation. In [8] the following three value creation systems are considered:

- The value chain,
- The value shop, and
- The value network.

These value creation system can be described as follows. The value chain is the classical production system studied by Porter [16] and is the type of value creation which exists in the production industry. Simply spoken, it consists of a chain of activities: logistics in representing the way raw material is obtained, production of the goods from raw material to finished product, and logistics out representing how the product reaches the market. The chain also contains several common activities. It is, of course, more to it than this, but this would suffice in order to understand the difference between the different value creation mechanisms. Competition between value chains will be the traditional one for market shares.

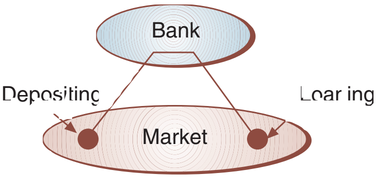


Figure 10 Value network: banking

The value shop is a problem solving organisations (consultants, architects and physicians are examples) where value is created by solving problems for clients. In telecommunications the system integrator is obviously a shop in this meaning. It only takes place in the production process when building an advanced solution. After that the system integrator withdraws from the process. The management of the continuous process is taken over by other roles.

The value network is represented by organisations acting as mediating agencies. They may connect users of the same kind as in the banking example in Figure 10. The bank connects people with money to deposit and people wanting to borrow money. In this case the network does not even possess the required goods, namely the money. The goods are in fact owned by the customers. The network may also consist of two types of co-operating banks: one having customers depositing money, and one having customers borrowing money. In this example the value of the customers of the co-operating bank is evident. This extreme example shows one important aspects of networks: the gain of co-operation. In this case the gain is obvious. In more complex cases it is not so obvious, and it may be difficult to assess the relative merits of competition and co-operation. It may even happen that two companies in network configuration may compete on certain products and co-operate on others. There are many indications that this may be so in telecommunica-

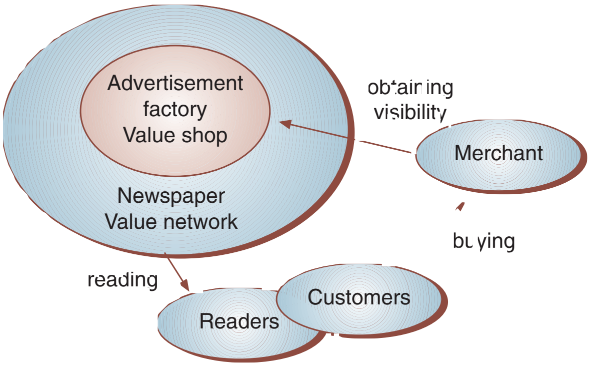


Figure 11 Combined network and shop: newspaper

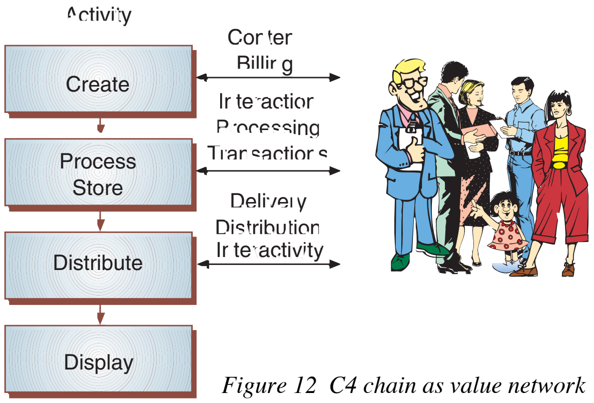


Figure 12 C4 chain as value network

tions. Telenor and Telia co-operate in connecting customers in Norway and Sweden, while they compete in each other's primary markets.

The network may also connect different types of customers. Figure 11 shows a more complex case where a newspaper acts as a network connecting merchants and their customers where some of the customers of the merchant are also customers of the newspaper. The merchant is also a customer of the newspaper by buying advertisements. The figure illustrates that different value creation organisations may interact. The newspaper is primarily a value network but contains a value shop producing the advertisements. The latter function could, of course, be undertaken by a separate agency. In this case the advertisement agency would have acted as a combined network and shop.

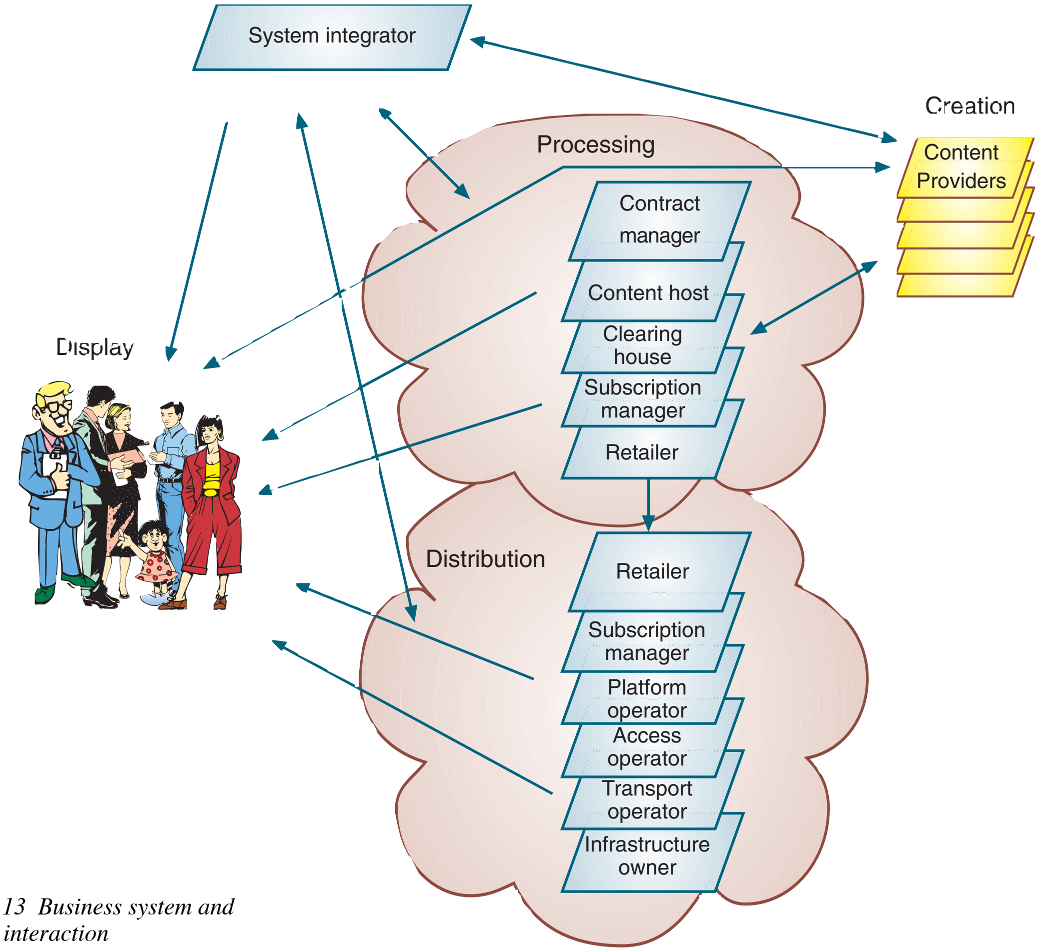


Figure 13 Business system and market interaction

Figure 12 shows the C4 chain as an interacting set of value networks representing content, processing, distribution and display. The main element here is that the providers of content, processing and distribution will deliver their partial products in parallel to the market. These deliveries are simultaneous and different, and none of them can be removed without collapsing the whole structure. Such a structure requires co-operation between companies in different layers. Companies in one layer may co-operate with companies which are competitors in another layer. A company distributing television signals may, for example, have as customers different production companies competing for viewers.

The business model for telecommunications is reproduced in Figure 13. Now it is put on top of the C4 model in order to identify in which part of the chain each role is located. The figure also illustrates that the different parts of the model have simultaneous market interactions. This applies not only to the 'C4 interactions' but also to interactions between the market and individual roles of the business model for telecommunications. This is illustrated in a few cases: content host and subscription manager in the processing domain, and the subscription manager and the access operator in the distribution domain. Note also that some roles are present in both the C4 activities distribution and processing. The system integrator has not been included in the C4 system. The reason is that one of the most important tasks of the system integrator is to create the complex value network required for each product. After this has been established, the system integrator is no longer a part of the business system delivering the product to the consumer. The type of value creating organisation of each role is shown in Figure 14.

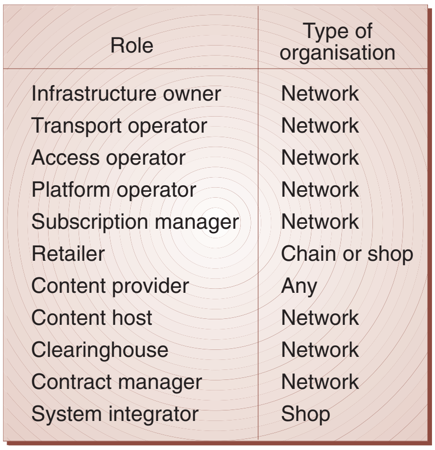
A final example is shown in Figure 15 also illustrating the interaction and co-operation between different value creating organisations. The products are comfort and economisation delivered by an energy producer, where the product will ensure the most economic use of energy for heating. The system consists of two chains, two shops and two networks interacting in a parallel delivery of the different components of the product. Two phases of the product are indicated because different companies are involved in the delivery in the two phases: the installation of the system (dotted lines) and the operation of the system (solid lines). The initiative for installing new equipment is taken by the system integrator and the actual installation is done by an installation company. Both may be regarded as shops from a value creating point of view. The system integrator will also choose which companies should deliver the different elements of the system. This is the first phase of the delivery of the product. In the next phase the installation company and the manufacturer of equipment will not take part in the delivery. The product is then delivered in parallel and simultaneously by the energy producer, a company handling the overall customer care and a company supporting the remote control. The system integrator may not participate at all in this phase. We have included it in order to indicate that it may deliver trust and guarantee in terms of brand name and other 'soft' products.

7 Complexity and Competition

Competition between the incumbent and newcomers in the new liberalised market takes place in several areas:

- In the marketplace where the aim is to get as many customers as possible;
- In the access area where the main issue is to install new accesses or to get hold of existing ones (leasing or local loop unbundling) in order to bind the customers;
- In the service area where the aim is to provide services over existing transport and access structures independently of the underlying network (point-of-sale, Internet);
- In entering into network roles like access, transport and platforms.

Competition in the marketplace takes place on price, on superiority of customer care, on products for special market segments both in the residential market and the business market, and on bundling of



Role	Type of organisation
Infrastructure owner	Network
Transport operator	Network
Access operator	Network
Platform operator	Network
Subscription manager	Network
Retailer	Chain or shop
Content provider	Any
Content host	Network
Clearinghouse	Network
Contract manager	Network
System integrator	Shop

Figure 14 Classification of roles and main value creating mechanisms

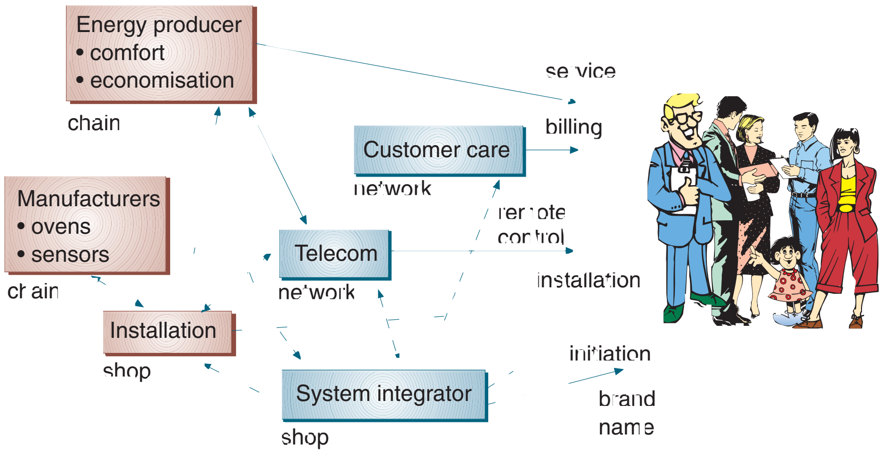


Figure 15 Interaction of shops and networks

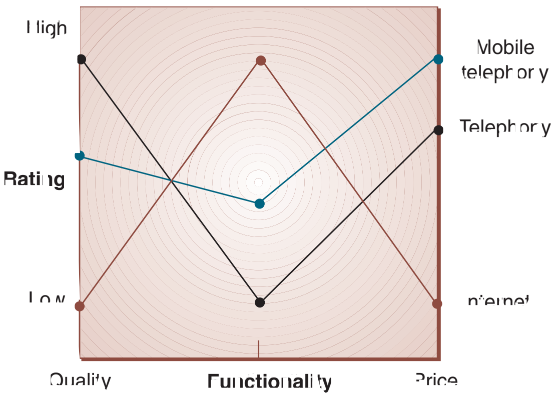


Figure 16 Parameterised competition

products. Today much of this competition is concerned with offering telephony services in the residential market. In this case, price will be the most important competition instrument since there is little room for advantages on other product parameters. In accordance with [2] this is mainly a price war type of competition. Provision of the Internet as an additional service is an added strategy where the focus is also on functionality (or quality as this parameter is called in [2]). This changes the competition strategy as we shall see.

Competition between mobile telephony and fixed network telephony is escalating. Mobile telephony has increased the overall market for telephone services by making the service more available. It is also replacing some of the need for fixed network telephony. This is also a type of competition where price is not the only parameter.

Competition between GSM operators is an excellent example of price war competition. Mobile services are special in one sense: they increase the availability of people on the move. This is an advantage of functionality. Initially this advantage of functionality was exploited and the prices were correspondingly high. The GSM system is such that it is very difficult to gain much from functionality: all parties offering the service can offer the same functionality at approximately the same cost. Therefore, much is not gained from a competition point of view where the aim is to capture parts of the market already held by the other operator. The other competitor must therefore create a bigger market. This can only be done by increasing it towards the end of the market where mobility is not regarded to be very important but of interest if the prices are sufficiently low. These market segments are then ignorant to the improved functionality provided by mobility. In order to increase the market in this way, the prices have to be lowered. This is an example where the market in terms of turnover is increasing slowly – and in extreme cases may become smaller, while the number of customers is increasing rapidly. The benefit for the user segments requiring high functionality is then a lower price as dictated by the new segment, while the benefit for the market requiring the lower price is more functionality: the mobile telephone may be used to convey such important messages as “I am now in the train, and what are we having for dinner?”. In the extreme case the prices for mobile communication and fixed network communications will become equal. In that case the two services have moved from initially being complementary to becoming fully substitutional: mobile systems extend the advantage of the cordless phone without geographic limits.

Such competition may thus be partly or wholly supplementary, complementary or substitutional, and may change over time. Figure 16 shows another example where fixed network telephony, mobile telephony and the Internet are compared using the three parameters (technical) quality, functionality (as perceived by the user), and price (the comparison is based on a method proposed in [17]). Even though all three networks may offer telephony, the Internet is the clear winner on functionality, and this is likely to be the key competitive parameter. Mobile telephony also scores on functionality but loses on price. Fixed network telephony cannot meet competition from the Internet by reducing the price because it cannot match the functionality. On the other hand, mobile services and the Internet may become supplements in the future by combining their functional merits.

The method outlined above may be a valuable method for analysis of competition between products and companies. As the market moves towards technological convergence and C4 convergence, there will be a number of such examples, some of which may be poorly understood and involve undesired possibilities for arbitrage. Some of these aspects may be uncovered by use of the method.

The second aspect is the importance of owning the access. In the GSM system the operator owns the access in terms of frequency licenses. The user is owned by not allowing him access to the competitor’s network by technical means, at least in areas where the networks overlap. The subscription contract may also be such that if the user changes operator within a certain time, he will have to forfeit some advantages. This situation is likely to change when the DCS1800 system is introduced since the roaming restrictions then are likely to disappear. The user needs a primary contact with an operator in order to manage his user profile. This management and the actual usage of the system will then be two different activities. This is again indicating the importance of the business system and the impact of the network economy described in Chapter 6. There will then be two different concerns binding a customer as a subscriber of user profile and making him use a specific access and platform to perform the telecommunications.

In fixed networks the access lines require the major part of the investments in networks. Most of this network today consists of copper wire, and it would involve a major expense for a competitor to replace it with modern technologies like optical fibres even though this would give advantages in terms of better functionality and a wider spectrum of services. One alternative for a competitor will then be to lease access lines from the incumbent or whoever owns them. Another alternative could be for the incumbent to offer the access lines to the competitors on conditions set by the regulatory bodies (local loop unbundling). The problem associated with such a decision is outlined in the next chapter.

The third possibility is that the competitor offers an overlay service to the customers independently of the infrastructure carrying it. The user will then be a customer of both the service provider and the owner of the underlying infrastructure. This is the way many Internet operators do it, and is in accordance with the co-operation strategy of value networks. Today the subscriptions for services and infrastructure may be independent. As this market expands, different subscription forms may emerge (see the discussion in Chapter 5 and 6 above). Co-operation may be beneficial for all parties in this structure.

The final case is where the competitor enters any of the roles defined in Chapter 6 in order to compete in single or combined niches. Here again the strategic questions are related to how value is created and the overall gain for all parties from an increased market caused by co-operation.

Competition – or co-operation – in the market is extremely complex. Above we have only pointed out some of the problems. Each of them requires advanced analysis tools like system dynamics, optimisation and game theory

8 Complexity of the Access: An Example

The above considerations are rather theoretical. Therefore, it will be useful to consider one example just to illustrate the complex analysis which is required only in order to understand the basic elements making up the system. A more comprehensive analysis will require several methods: dynamic performance of the business system using cybernetic and econometric methods, stochastic analysis and determination of the impact of uncertainty [18], game theoretic analyses determining the merits of co-operation versus competition in certain cases, and other methods such as analysis of competencies [19]. Each role in the business model of Chapter 6 needs to be modelled separately. Here we will give an analysis of the access role and identify which parameters that need to be put into the more comprehensive cybernetic, stochastic and game theoretical models.

Access operation is chosen because it is one of the most complex roles of telecommunications. There are several reasons for this:

- The control over the access and the ownership of the users are regarded to be synonyms leading to various strategies such as building separate access infrastructures, leasing access lines, use of prefixes and other discriminating means in order to accommodate several operators on the same access infrastructure, bulk sales of access capacity, and local loop unbundling where the conditions for obtaining access infrastructure are part of the sector specific regulation of telecommunications.
- Some access technologies are becoming substitutes for one another. Substitution may come as a result of the development of a certain technology such as optical fibres substituting copper wires, or because different technologies may offer similar service capabilities such as radio access substituting copper wires.
- Some technologies are supplementary to one another; that is, one technology adds capabilities to another technology enabling enhanced offering of services. In this way an Internet access may be a supplement to a copper wire access (though it may be a substitute for the telephone *service*) enhancing its basic features.
- Some technologies are complementary, that is, they offer capabilities which neither replace nor enhance each other. One example is that copper wire access and radio broadcasting are complementary access technologies.

One problem is that the categories are seldom pure. Access types may, for example, be substitutes for each other in some cases and complements in other situations. Copper wire access is complementary to some cable TV systems but competitive to others.

Another problem which complicates the issue further is that evolution of technologies, markets and prices may change the category to which a given set of accesses belongs or even enable new technologies to be developed such as the 'electricity modem' on the local electricity network. GSM and copper access were initially complementary, that is, GSM was mainly used by customers for whom mobility was important and did not affect their usage pattern of fixed network telephony. The price level is now becoming such that GSM replaces copper wire access offering more functionality. DECT[8], UMTS[9] and other mobile technologies may do the same.

For some of the technologies only one operator can be present at a time. On a single copper wire or fibre there can only be one access operator present. In other technologies there may be several operators offering the same system at the same time. Radio systems are examples of these. In both cases each operator needs an access infrastructure such as cables and base stations. However, the alternative to the first is that two or more systems (wires or fibres), one for each competitor, are built to each customer, then leaving the choice of operator entirely with the user. This alternative is, of course, economically not feasible except in a few cases (existing cable TV systems and copper wire telephony).

A final problem is that some accesses are unique for each user (wires or fibres) while others (radio) are shared by many users. This also gives different economies to different access techniques. In some of these systems the cost per user may on average be the same independently of the number of users (copper wire or fibre). In other cases, the access cost is reduced as the number of customers increases (mobile telephony), or the access cost per user may increase if the number of users increases (in mobile systems where the cell structure must be rearranged).

With all these possibilities, access becomes a multidimensional problem which is very difficult, if at all possible, to attack by use of conventional means. The dimensions are at least the following:

- Replacement capability as discussed above;
- Access type segmentation;
- Basic infrastructure;
- Geographic segmentation;
- Market segmentation;
- Service segmentation;
- Evolution and innovation;
- Functionality and bundling;
- National versus international access;
- Pricing method.

This gives us a ten-dimensional problem where, for example, unbundling must be understood in all dimensions. The first item – the replacement capability – was discussed above. The others will be considered briefly below. There may be

[8] *DECT is the digital European cordless telecommunications system.*

[9] *UMTS is an abbreviation for Universal Mobile Telecommunications System which in most respects can substitute the ISDN.*

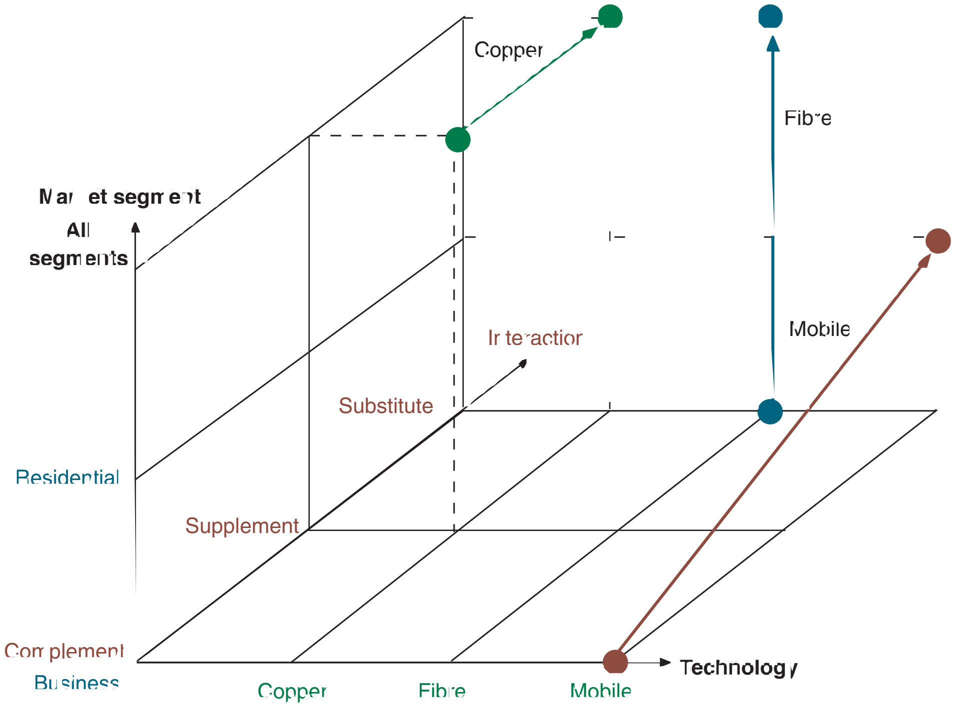


Figure 17 Example of 'product space'

several more parameters which should have been included.

A given system will occupy one point in this ten-dimensional space and follow one trajectory as the system evolves in the future. Another system will occupy a different point now and follow a different trajectory into the future. Two systems will most likely not be identical and occupy the same points of this space. This makes it extremely difficult to identify when rules defined for competition are met or violated. An example is shown in Figure 17. We have included only the three parameters technology, market segment and interaction, and shown a possible evolution at two instants. Here we have assumed that copper wire access is moving from being a supplement in all markets to becoming a substitute (that is, one of many techniques) in all markets, optical fibre access is shifting from being a substitute in the business market to also becoming a substitute in all markets, and mobile access is moving from being a complement in the business market to becoming a substitute in the residential market. The question is: when and where are these technologies competing?

8.1 Access Types

The basic access types are (omitting disappearing technologies like NMT):

- Copper wires capable of analogue transmission or ISDN. ADSL[10] and similar technologies are supplements offering more bandwidth. The basic technology exists everywhere and several supplements are possible, but their market potential depends mainly on price. The bandwidth that can be used on wires is therefore increasing, which may imply that existing cables have a much longer lifetime than expected.
- Optical fibres offering 'unlimited' bandwidth to the user. The technology exists in the business segment and is coming in the residential market as cable TV networks are enhanced and copper wires are replaced in areas where cheap infrastructure for laying fibres exists.
- GSM in the 900 MHz frequency range allows roaming between almost all countries, and is besides copper wire the most used access technology in the world.
- DECT has the same capabilities as GSM and may replace it. Or it may offer local mobility and thus not replace GSM but take traffic away from that system and fixed networks. However, it is more likely that the system may be replaced by other mobile technologies since the concept is becoming obsolete before it is installed.
- TETRA[11] is primarily a system for private trunked radio but can be used for public access and compete with the GSM system, at least in cities. The technology is based on the GSM technology.
- DCS1800 is a GSM technology in the 1800 MHz band offering more bandwidth. This technology is now being implemented in several countries, and the manufacturers are making dual mode mobile stations operating both in the 900 MHz band and the 1800 MHz band, and several regulating authorities are requiring roaming to be possible between systems in the two bands also when the operators compete in the same geographical area.

[10] *Asynchronous Digital Subscriber Line.*

[11] *TETRA is an abbreviation for Trans-European Trunked Radio Access.*

- UMTS will probably replace GSM as the next generation land mobile system, or at least this is the intention behind specifying the technology.
- Combinations of terrestrial radio solutions are possible. Examples are GSM/DECT and GSM/DCS1800.
- Cable TV networks offering one-way broadcast services is also an access technology. This technology may also become a combined substitute and supplement for some applications on the copper wire. The systems will have the capability of offering broadband Internet services to the user and combining it with narrowband technology for the return channel in co-operation with another operator. This will reduce the time for downloading information from the Internet considerably.
- Cable TV networks can also be enhanced for two-way services as has been done in several countries.
- One-way radio broadcast networks (terrestrial or satellite) may be enhanced in the same way as cable TV networks.
- Geostationary satellites can offer broadband transmission one-way and combine it with a narrowband system for the return direction, for example, offering Internet services.
- Mobile geostationary satellite systems such as the INMARSAT system.
- Low orbit satellite systems such as IRIDIUM and Teledisc.
- ‘Eelectricity modems’ enabling 1 or 2 Mbit/s over electricity supply systems in the residential market.
- ‘Balloon satellites’ where cheap satellite repeaters are placed in balloons for medium area coverage for both two-way systems and broadcast systems.
- Broadband two-way land mobile systems (BMS) are being developed offering capacity comparable to that of optical fibres.

Note that most of these technologies were put into service or developed during the last five years, and that some of them are still being developed. Note also that in many cases two accesses are required in order to establish the pertinent connection between the user and the service provider. Examples are

- Internet over any of the above basic technologies;
- Modem access to data networks over wire, radio or fibre;
- Frame relay over wire, cable and fibre;
- Multimedia access in the future over any of the above.

Some technologies have been developed but are still not available because there has not been a big enough market for them. This may change radically as the liberalised market is maturing and the control of the access becomes an increasingly important issue for the newcomers.

8.2 Basic Infrastructure

The existence of a basic infrastructure and the cost of building a new infrastructure determine the threshold for new entrants to establish themselves in the access market. The access infrastructure is the most expensive part of the network since the possibility of utilising traffic concentration and dividing the investment costs on many accesses is small. Infrastructure elements which may be used are:

- Ducts, pipes and tubes like sewers and gas pipes, or ducts deliberately made for various use in the future (along roads, railways and so on);
- Poles for road lighting and electricity supply suitable for mounting radio antennae for mobile communications systems like TETRA and DCS1800, and for placing sensors like microchip video cameras;
- Electricity supply systems that can be used for roll out of optical fibres, in particular to the industry;
- Unbundling the copper line access of the incumbent where a competitor is buying or renting the existing systems;
- Utilisation of the same duct for several purposes when new areas are built out or some infrastructure is maintained or replaced in an area. This is particularly important for electricity companies if they also offer telecommunications where they may distribute the cost on two (or more) products (energy and telecommunications);
- Radio frequencies, where in some cases the obstacle is licensing of frequency usage making this to a force which is not entirely market driven. On the other hand, the availability of frequency licenses will probably become one of the most important competitive assets in the future.

A new entrant can utilise a combination of infrastructures in order to find the minimum cost for building a new system. Local loop unbundling and frequency licensing are regulatory matters which will have a major impact on the competition in the market. Local loop unbundling is particularly important since it may open for arbitrage (buying capacity and portioning it into smaller units, for example, dividing a 64 kbit/s channel into four 16 kbit/s channels by use of voice coding techniques), it may leave uncertainties with regard to responsibilities and liabilities, and it may slow the evolution of access technologies since it removes the need for this type of innovations.

8.3 Geographic Segmentation

The residential market and the business market are rather concentrated in many countries, and a new entrant will enter selectively in areas which will offer the best business opportunities. There will certainly also be areas where no-one will enter. The geographic segmentation will be of particular importance for the incumbent because in most countries dominance is defined in terms of market share (for example dominance may mean that the total market share is above 25 %). This may mean that a company may hold most of the subscribers in a country and being dominant without earning money since the competitor has taken the most lucrative niches offering good margins.

Several values may be given to the geographic segmentation variable:

- The country is regarded as a single segment with regard to market share and competition and does not take into account geographic concentration of people and businesses;
- A distinction is made between cities and rural areas in the rules governing market shares;
- The country is divided into areas such as counties or municipalities and the market share of competitors may be computed for each segment of the grid.

It may then be possible to put different weight on the segments and competition within each segment:

- Each competitor should have the same capability as the incumbent in every segment of the grid;
- All competitors combined should have the same capability as the incumbent, but there may be different competitors in different geographical segments;
- There may be areas where there are natural monopolies. These areas are not considered with regard to market share;
- Only the market share for the core businesses of telecommunications (telephony, transport systems and copper wire access) is included in the evaluation but not other technologies. In such regulations it is, for example, ignored that telephony over copper wires is competing with mobile systems;
- It is recognised that different technologies compete, and the definition of competition and market share is widened to take this into account.

The different strategies will define different competition conditions, and it may lead to different strategies for co-operation between the various actors in the marketplace.

8.4 Market Segmentation

The telecommunications market is complex and consists of several segments. First, we have the important differentiation between the residential market and the business market. Then within each of these there are several segments depending on total usage, usage patterns and service requirements.

Within the business market one important factor is the size of the customer and the amount of expenses the company uses on telecommunications. Another not so obvious differentiating factor is whether the customer produces much outgoing traffic or much incoming traffic. It is easy to measure the former and determine how much the operator earns from that user. However, the second type of customer may create much income from the users calling this customer or from interconnect fees for terminating traffic from other operators. In the business structure described above (Figure 9), the income from traffic may thus come in one role while there is another role being responsible for the customer terminating the traffic. This situation will be typical for premium rate services and for Internet services. Travel agencies and taxis are examples of such asymmetric customers.

Competitors will search for market segments offering much income and healthy margins, and ignore others. Therefore, direct competition may not occur in certain market segments, while other segments will have many competitors. The market in one way resembles the car market where there is no competition between Skoda and Mercedes because they attack different market segments, while there are other manufacturers offering products to several segments competing with both of them (for example the Ford company and most Japanese companies). The competition is so new in telecommunications that this type of business has not yet developed.

The market segment variable may then be characterised by the following variables:

- Residential market with several subsegments where each segment is characterised by usage volume and usage patterns;
- Business market divided into categories based on the size and type of the business and with several subsegments based on traffic direction such as high outgoing traffic, high incoming traffic, or both.

8.5 Service Segmentation

Note that here we refer only to services offered by the access and not to services offered by a platform to which the access is connected. These cases must be kept distinct because the access deals only with the way in which value added services are offered and not how they are produced. I mentioned above that the Internet will require an additional access on top of the access offered by copper wires, fibres and radio. In this case one access technology is concerned with how the user is connected to the network. The other access (to Internet) ensures that the interactivity between the user and the network is maintained.

Some technologies may only offer a limited set of the access services listed below. Other technologies may offer a larger or more sophisticated range of services.

Noting that the applications in terms of telecommunications services as we are used to them are not offered by the access itself but by the platform to which the access is connecting the customer, this variable may take on a large number of different values. The access may then offer capabilities as follows:

- Narrow bandwidth offered to residential users with standard supplementary features;
- Narrow bandwidth offered to businesses;
- ISDN (two channels and more flexibility) offered to residential users;
- Extended ISDN offered to businesses;
- Broad bandwidth offered to businesses and residential users;
- Dark fibre offered to businesses;
- Asymmetric solutions where there is a broadband system from the network to the user, while the opposite direction is served by a narrowband system;
- Supplementing (and possibly more expensive) technologies like ADSL.

8.6 Evolution and Innovation

The competitive environment must be such that it favours evolution and innovation. This is probably the most difficult issue to regulate for several reasons:

- Unbundling of the copper wire may give all competitors equal access to the users so that there is no competitive force favouring evolution of alternative and better solutions. Under such regulations, will fibre to the home come at all?
- How should the operator in such a regime finance a joint evolution of the technology?
- How should an operator enhance the network if the use of it is sold to another operator? Who should pay for such innovation or will there be any sharing of risks? Who is responsible for faults?
- Access from individual customers are often multiplexed on the same transmission system in order to reduce the cost per access. This increases the problems of responsibility and liability if something goes wrong.

This problem is associated with the problem of defining the extent of the access network, also noting that this is technology dependent and subject to evolution. The fixed network has evolved towards fewer exchanges (10 % of what we had 20 years ago) but with a large number of concentrators and remotely controlled switching units with limited service capabilities. This evolution may continue depending on the general evolution of transmission systems. On the other hand, a land mobile system contains fewer switches and has a much simpler transmission structure. It requires a large number of base stations so that the access is not necessarily cheaper. This makes it difficult to compare them and to see where the evolution will take them with regard to market penetration and competition in the future.

It is also reasonable to assume that it is more likely that new access technologies will be developed if the competitors are forced into a situation where they have to be innovative rather than copying old and obsolete technologies. This may give the incumbent some initial advantage but technological evolution may take that away soon (the ‘electricity modem’ may be such a leap).

8.7 Functionality and Bundling

Market penetration is not only dependent on the cost of the access. This is the case if the product is a primary product. As we defined it above, telephony in the residential market is such a product. We also saw above that it is possible to bundle it with other products such as cable TV, energy and the Internet. This may change the value of the telephony product as perceived by the user. It may also change the price of the product, either by cross subsidising or by lower cost, because some cost elements can be distributed over more products in the bundle. The intelligent building is such a concept where the cost is divided among products like telephony, energy delivery and surveillance. In this case, it may be argued that the bundling is entirely within telecommunications since the user pays for the remote control and monitoring facilities.

This part of the problem is very complex and the possibility and capability of bundling may change over time and hence become an evolutionary parameter.

A less visible bundling comes from operators offering Internet telephony. This product is poorer with regard to quality and reliability than ordinary telephony, but what makes it attractive is not only, as is the general feeling, that it may be cheaper but that it offers more functionality being associated with text and graphical services. The service may be triggered by clicking hypertext pointers in the same way as for accessing the Internet from locally stored information. This service is offered from encyclopaedia on CD-ROM for access to supplementary information or updates: buying the CD-ROM, you also buy the access to information.

All of the technologies listed above should be characterised with regard to their differentiating capabilities, both alone and in different bundles.

8.8 National versus International Access

We always assume that competition is national and do our assessment based on that. However, low orbit satellite systems (IRIDIUM, Teledisc) may be important threats to national access operators. Such systems will produce relatively little delay in the speech channels so that they may offer a quality of service not significantly different from that of the fixed network.

Other types of competition will be met from other satellite operators. Internet over broadcast satellites is again a good example. This may change the telecommunications business completely because geographically unbound operators may take over part of the business. This is not unlike the way in which long distance satellite communications removed the lucrative businesses of transit traffic.

8.9 Pricing Method

The access competition may be obscured because of different pricing mechanisms. We have already seen one such problem related to bundling where cost is divided among several products. Other possibilities are:

- Access financed by advertisements;
- Access financed as part of house rental;
- Access financed by value added service such as electronic commerce.

9 Conclusions

The most important strategic competency for companies within telecommunications is to understand the complexity of the business and to act in accordance with this understanding. The source of much of this complexity is new, and the business is changing for many reasons:

- The telecommunications technology has been through a rapid change during the last decade. Not only has the focus changed from the primary concern of being competent in transmission to putting content in the front seat, but also to identify and master applications which can be made from general purpose processing. This part of the evolution is still going on but a new component is entering: technology now moves towards cheap sensors and nanotechnologies opening for new applications in monitoring the environment, people and machines;
- The telecommunications industry developed the transmission technology but has not been the driving force in the evolution which has taken place afterwards. The data industry, the content industry and the academia have been the main driving forces in this development. The microelectronics industry may enter next in the development of the new sensor technology;
- The dependability of society on telecommunications is increasing. This applies to all parts of society and to more and more of its activities, and the new technologies will make this dependability even more critical. Dependability and telecommunications should become a new area of academic studies;
- The organisation of the business is also changing rapidly. This is partly because of liberalisation of the market, but the technological evolution has probably been just as important for this change. The new technologies will bring this organisational change even further. Telecommunications may then become a commodity supporting other applications and be integrated in other businesses leading to an advanced form of C4 convergence. Making money in such an environment depends very much on understanding how value is created in the organisation [8];
- The products are also taking on new forms, changing from being mainly

two-party services to accessing content and information. This evolution is in many ways a result of the growth of the World Wide Web and its supporting technologies. The advance of the sensor technology may again change this picture, and that even before the previous technologies and businesses have matured and stabilised;

- This indicates that mastering advanced methods in strategic planning may become key factors in the future positioning in the market. These methods should build on game theory, cybernetics (both feed back and feed forward methods), optimisation and spatiotemporal evolution theory ([1], [20], [21]).

There is no other industry where similar changes have occurred, so that there is not much to learn from experience and analyses of other businesses. The strategy for telecommunications must therefore be built from scratch.

Most of the development of strategy as a scientific subject has taken place since 1970. The full recognition of the subject came around 1980 when the Strategic Management Journal was founded. Still strategy is an issue which is not fully recognised. The best methods are still not in common use, they are simply too complex. In telecommunications we probably need them. One mind opening survey is contained in [22], showing how complex strategies may be.

Jan A. Audestad (56) is Senior Advisor for the Corporate Management of Telenor. He is also Adjunct Professor of telematics at the Norwegian University of Science and Technology.

e-mail: Jan-Arild.Audestad@s.hk.telenor.no

References

1 Dendrinos, D S, Sonis, M. *Chaos and socio-spatial dynamics*. Springer-Verlag, 1990.

2 D'Aveni, R A. *Hypercompetitive Rivalries*. The Free Press, 1995.

3 Saffo, P. *Closing session of ITS 98*, Stockholm 21 – 24 June, 1998.

4 Gemperlein, J. *Paul Saffo, tech forecaster*. San Jose Mercury News. Available on World Wide Web under the title Q and A with Paul Saffo.

5 Regis, E. *Nano!* Bantam Books, 1997.

6 Cochrane, P, Heatley, D J T. *Modelling future telecommunications systems*. Chapman & Hall, 1996.

7 Audestad, J A. Challenges in a liberalised telecommunications market. *OR98 Conference Proceedings,* Zürich, August/September, 1998.

8 Stabell, C B, Fjeldstad, Ø D. Configuring value for competitive advantage : on chains, shops, and networks. *Strategic Management Journal,* 19, 413–437, 1998.

9 ISO/IEC/ITU. *Reference Model of Open Distributed Systems Part 2 : Descriptive Model*. ISO Document 10746-2/ITU-T Recommendation X. 902.

10 Audestad, J A. Strategy for enterprise specifications. *Telektronikk,* 94, (3/4), 1998, 24–32. (This issue.)

11 Hagel III, J, Armstrong, A G. *net.gain*. Harvard Business School Press, 1997.

12 Cairncross, F. *The death of distance*. Harvard Business School Press, 1997.

13 Baldwin, T F, McVoy, D S, Steinfeld, C. *Convergence : Integrating media, information & communication*. Sage Publications, 1996.

14 Handegård, T, Kristiansen, L. The TINA architecture. *Telektronikk,* 94, (1), 95–106, 1998.

15 TINA Consortium. *TINA Business Model and Reference Points, Version 4.0,* May 20, 1997.

16 Porter, M. *Competitive advantage : creating and sustaining superior performance*. Free Press, 1985.

17 Kim, W C, Mauborgne, R. Value innovation : creating and sustaining superior performance. *Harvard Business Review,* Jan.-Feb., 1997.

18 Wallace, S W. The role of uncertainty in strategic planning. *Telektronikk,* 94, (3/4), 1998, 21–23. (This issue.)

19 Audestad, J A, Kjæreng, A, Mahieu, L. An object oriented model of the telecommunications business. *Telektronikk,* 94, (3/4), 1998, 54–61. (This issue.)

20 Rumelt, R P, Schendel, D E, Treece, D J. (eds.). *Fundamental issues in strategy*. Harvard Business School Press, 1994.

21 Anderson, P W, Arrow, K J, Pines, D. *Economy as an evolving complex system*. Santa Fe Institute Studies in the Science of Complexity, Volume V, Addison-Wesley, 1988.

22 Kelly, K. *Out of control : the rise of neo-biological civilization*. Addison-Wesly, 1994.

23 Korzeniowski, P. *Middleware : achieving open systems for the enterprise*. Computer Technology Research Corp., 1997.

The role of uncertainty in strategic planning

STEIN W. WALLACE

1 Introduction

Strategic planning and uncertainty[1] are twins. The first makes no sense without the latter. It is the very existence of uncertainty that forces us to produce strategic plans. Had we known the future, strategy would have been without interest, even without meaning.

This very obvious observation leads to some conclusions that in my view ought to be as obvious. If uncertainty is an important part of the problem, it is reasonable to think that it must also be an important part of the methodology used in planning. By methodology I now refer to both quantitative models and qualitative reasoning. And even more importantly, it is not at all unimportant how the uncertainty is entered into the planning process.

Before continuing, let us note in passing that although we are now discussing strategic planning, most, if not all, of our conclusions also apply to tactic planning. In my view there are no deterministic planning problems. There are certainly planning problems that are well studied by deterministic methods, but how do we know that the one we are looking at is in that class? The rest of this note will indicate what may go wrong if we rely on deterministic tools in what is truly a stochastic problem.

2 A small example[2]

Assume that a provider of telephone services, operating under competition, is faced with the problem of how to develop its infrastructure in a residential area under development. The infrastructure must be built while the residential area is under construction, and before any customer has decided which telephone company to use. The area is planned for 10,000 housing units. In monopoly times this meant almost surely 10,000 customers. But now, under competition, things are different. When the market division of the company is asked how many customers the company can expect, the reply is that they expect 6,000, but, in fact, they could get any number of customers between zero and 10,000. This would depend on their own strategies, as well as those of the competitors. However hard pressed, the market division refuses to be more specific than this. The management therefore turns to its planning division with this information and asks what would be the correct decision in the different cases. The planning division replies by producing a figure, as the one in Figure 1.

From Figure 1, management sees that whatever number of customers it obtains, it will be optimal with a fixed network. After all, there exists no possible future (i.e. customer base) where mobile is better than fixed. It is therefore decided that they will build an infrastructure for a fixed network. A further analysis is performed to determine the size of the fixed installation. The approach is as follows. Each of the ten solutions obtained to produce the columns for fixed networks in Figure 1 is tested towards all possible futures to find which one has the best expected performance.[3] For example, we check how the solution for 3,000 customers behaves if in fact 7,000 customers show up. This way the best of the fixed network solutions is found.

But then a stubborn member of the staff insists on checking also the expected performance of the mobile solution, despite the fact that no possible customer base exists where a mobile solution is best. And, lo and behold, its expected performance is substantially better. What is going on? Figure 2 *illustrates* the point. While the mobile solution is never best when the number of customers is known, it is at times much better than any of the fixed solutions whenever the number of customers is different from the best possible value.

3 What have we learned?

There are many things to be learned from such a simple example. Because, indeed, the example is very simple. So simple, that it is tempting to think that the logical errors made above are unreasonable. Also, it may be tempting to start discussing the model rather than the conclusions, based on the feeling that the problems are caused by the simplicity of the model. That, however, is not the

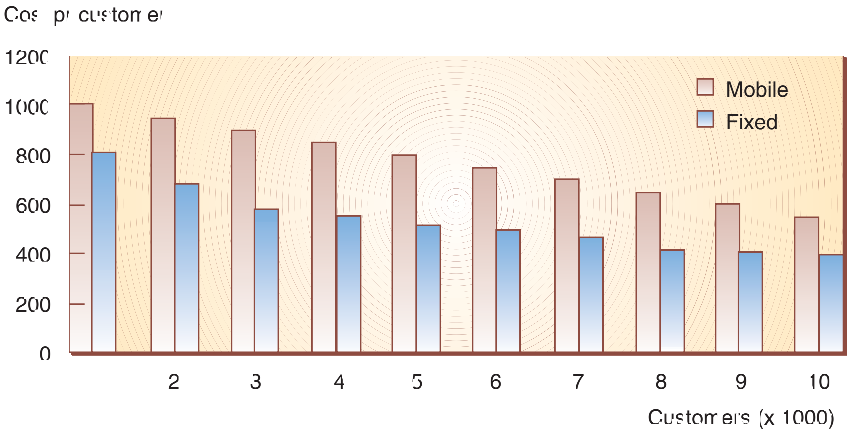


Figure 1 Cost estimates from the planning division. For each possible customer base, the cheapest mobile and the cheapest fixed solution is given

[1] *In this short note we do not make any distinctions between stochasticity, uncertainty, randomness or risk. In some contexts that may be important; here it is the relationship between stochastic and deterministic that is in focus.*

[2] *This example was created in co-operation with Nils-Jacob Berland.*

[3] *Of course, we could also maximize expected utility, or even perform a worst case analysis. The arguments to follow will function also for these cases.*

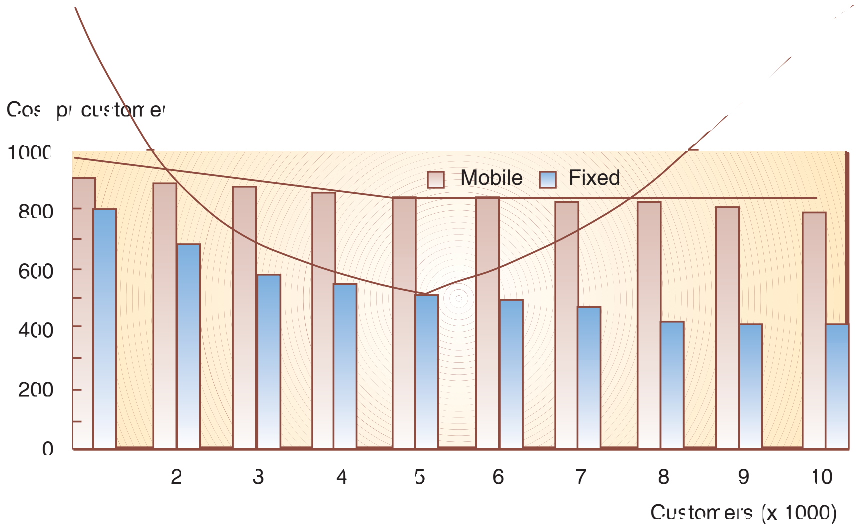


Figure 2 Principal illustration of how a single solution behaves if the customer base is different from the one corresponding to that solution

case. The erroneous thinking used above is what we find in almost all textbooks, and many software packages advertise that they facilitate exactly this type of approach.

What went wrong in our argument, and what is indeed a common error, is to base decisions on the following assumption. *If we find the optimal decision (plan, strategy) for* all possible futures, *and all these decisions share some property, we feel confident that this property must be part of the solution we are seeking, namely the one which is best in an over-all (expected) sense*. We think that by studying *scenario solutions* we learn something useful about the problem under investigation. As we saw above, that is not at all a generally correct observation. For absolutely all possible futures the best was to install a fixed network, and that was *exactly* what we did not want.

Often, the starting point of an analysis is a deterministic model where all random variables are replaced by their means (or some other representative). The parameters are often not explicitly interpreted as means, but implicitly that is what they are. This will produce some decision (strategy). Since the decision maker knows that the parameters used may be incorrect, he then resorts to some parametric analysis to see if the first solution is stable. If he finds that the solution remains optimal (or aspects of the solution remain intact) for many possible parameters, he feels confident that the first solution was good. What we saw in the example is that this is false. Not only does the parametric analysis not yield what it promises, it is directly counter-productive. Before the analysis the decision-maker was uncertain about the quality of the solution found from the deterministic model, afterwards he feels more confident, despite the fact that it is incorrect. The analysis has fooled him. *We see that sensitivity analysis simply is not a way to understand and analyze stochastic decision problems*. A more detailed discussion of this fact can be found in Wallace [2].

3.1 Optimal in hindsight

For the example, note that in hindsight, that is, after the fact, one of the scenarios will have happened. Hence, one of the scenario solutions will be optimal. Therefore, in hindsight, it will appear that a fixed network solution is best. We know for sure that the solution we advocated, namely a mobile solution, will not look extremely good. This is important to note, as this is probably a major reason for imagining that looking at scenario solutions is the right thing to do. After all, one of them will certainly be optimal after the fact. And still, we might be totally mislead. The problem is of course that if we pick the wrong scenario solution, it may turn out to be very bad.

3.2 Evaluating decision makers

As we have seen, decisions which are optimal in terms of expected behavior are normally never optimal in hindsight. Hence, decision-makers that follow our advice, and which are the best decision-makers in the long run, will never make the really outstanding decisions. This is important to remember when decision-makers are evaluated. Do we value the decision-maker who occasionally makes marvelous decisions, or the ones who are good in the long run? So care must be taken with respect to hindsight, as the IQ of hindsight is very high.

4 So what are scenarios?

We are of course free to define scenarios any way we want. But for scenarios to be useful in a context of decision-making, they must have one crucial property. *A scenario must be independent of the actions of the decision-maker*. A scenario therefore describes the (random) environment facing the decision-maker. Hence, statements such as “it will rain tomorrow”, “EU regulation will allow competition” and “demand will increase” are most likely scenarios, whereas “we will choose to cooperate with Telia” is not a scenario (unless the decision maker views decisions in his own company as random events he cannot affect or control.)

Looking at our example, we may wonder if the number of customers is a good random variable, i.e. have we defined good scenarios? If we can affect the number of customers for example by marketing and price schemes, our scenarios are not particularly well defined, as they are not independent of our actions. Rather, we should have tried to define scenarios in terms of general trends among customers, governmental regulations, and possibly strategies of competitors. Scenarios may be defined in terms of such as the general popularity of mobile solutions, whether or not our competitors will develop a fixed network, and to what extent we will be allowed to compete. Decisions may then be such as marketing, market research, and, of course, our choice between a mobile and a fixed network infrastructure.

It is important to note what we argued against in the first section. It was not the use of scenarios as such, nor was it the fact that the scenarios possibly were not too well defined. The point was that scenarios cannot be used to *find* good solutions, since good solutions rarely are solutions to scenario problems. But scenarios are appropriate for *evaluating* possible solutions. That is, given a solution (strategy), and a set of scenarios describing the possible futures, we can correctly find the expected performance of that solution. This is a process we would often refer to as simulation.[4] So scenarios are indeed important for decision making under uncertainty, as long as they are appropriately used.

5 Scenario aggregation

Methodologies for using scenarios to find optimal solutions to planning problems do exist. Having understood that the difficulty with scenario analysis is its inability to *find* optimal solutions, and not its ability to evaluate potential solutions, we realize that what we need is a more advanced setting for the optimization. (And this applies also to decision-makers who make decisions based on qualitative analysis.) Verbally, what we need to do is to set up a model that searches for a solution that is optimal in the expected sense, rather than being optimal in one single scenario.

Let s be a scenario (possible future), $p(s)$ the probability that s will occur, and S the set of all scenarios. Assume our problem is to maximize the expected profit $E_S f(s,x)$ with respect to x. Here E_S means to take the expectation with respect to the scenarios in S, and $f(s,x)$ is the profit if we make decision x and scenario s happens. So the goal is to find the x which has the best expected performance. The scenario approach, on the other hand, is to solve for one s at a time

$$\max_x f(s,x) \text{ to obtain } x_s .$$

Then, to find which of these solutions behave the best, we solve

$$\max_{\{x_s \text{ for } s\in S\}} \sum_{s'\in S} p(s')f(s',x_s) \text{ to obtain } \hat{x}.$$

What we have found out earlier is that $\hat{x}$ can be a rather bad solution, despite being the best of the scenario solutions. Instead we should solve

$$\max_x \sum_{s\in S} p(s)f(s,x) \text{ to obtain } x_* . \qquad (1)$$

The latter problem is of course much harder to solve, as it is much bigger. It is also harder to grasp mentally if one resorts to qualitative approaches. Because also in the qualitative setting, a scenario based approach will lead in the wrong direction.

It is not the purpose of this note to discuss the issues of formulating and solving stochastic programs, such as (1), but rather to point out that such complex formulations may be necessary, and that procedures are available. For a detailed look at the field, see for example the textbook by Kall and Wallace [1].

5.1 Decision trees

Many are used to applying decision trees for studying decision-making under uncertainty. A decision tree is a tree that has two types of nodes, one for possible decisions, and one for realizations of random variables. Decision trees are logically correct, and yield optimal solutions. Note however, that for such trees, there must be a small and finite number of possible decisions. Hence, the potential actions to be taken are input to the model, and the output is just information about which one is best. Hence, decision trees do not help us find potential solutions, and in that sense suffer from partly the same problems as scenario analysis. But if the number of potential solutions is indeed clearly finite, decisions trees are appropriate.

6 Conclusions

The main message of this short note has been that the use of scenarios for finding potential solutions (strategies) for a decision problem under uncertainty (be it strategic or tactic) can lead to arbitrarily bad solutions. Scenario solutions do not offer information about the structure of the solution that is best in expectation. Hence, other, and more complex methods are needed, and these do indeed exist.

Scenarios are on the other hand appropriate for evaluating existing potential solutions. Technically speaking, this is what we normally refer to as simulation.

References

1 Kall, P, Wallace, S W. *Stochastic programming*. Chichester, Wiley, 1994.

2 Wallace, S W. Decision making under uncertainty : is sensitivity analysis of any use? *Operations Research*. To appear.

Stein W. Wallace (42) has been Prof. of Operations Research at NTNU since 1990 after spending 7 years as Senior Scientist at Chr. Michelsen Institute, Bergen, and 3 years at Haugesund Maritime College. Throughout, his focus has been on theoretical and applied aspects of decision making under uncertainty, with applications in fisheries management, petroleum field development, portfolio management and telecommunications.

e-mail: sww@iot.ntnu.no

[4] *Although the term simulation is used with many different meanings, and is often not very well defined.*

Strategy for Enterprise Specifications

JAN A. AUDESTAD

1 Why Open Distributed Processing?

Telecommunication is a technological area and the technology must be the basis for the definition of the strategy for such systems. Telecommunication systems are also distributed systems by their very nature. It is, however, new that distribution has been taken into account in computer systems in general. To a great extent such systems have been modelled without clear considerations on distribution. During the last few years, however, distributed processing for general computer applications has been developed jointly by the ISO, IEC and ITU[1] ([1], [2]). The idea is that this specification should become the basis for all distributed computer systems. The first commercial products are available now (CORBA[2]).

The specification is called Open Distributed Processing, and is abbreviated ODP. We will use this abbreviation throughout. The purpose of this essay is to use these specifications in order to demonstrate how this specification can be used to formulate a strategy of telecommunications systems in general. In particular, we will concentrate on the methods used for formulating the requirement specification of a system, in ODP referred to as the enterprise viewpoint specification.

> *Note that every system can – and should – be specified at the enterprise viewpoint using the principles presented here. These principles are general and independent of whether the system in question is an ODP system or another type of system.*

Since 1993 work has been progressing on a new architecture for telecommunication based on this model. The work has been a co-operation within the TINA[3] Consortium owned jointly by telecommunications operators and manufacturers of computers, software and telecommunications equipment. There are several concerns behind this development:

- Telecommunications and information technologies are converging so that there are several commercial and technological advantages of developing a common platform for such systems;
- The TINA platform should replace the intelligent network nodes of the existing networks with more modern equipment with better scalability performance and more evolutionary potential;
- The TINA platform should convert the telecommunications networks into systems with the same flexibility as the World Wide Web while meeting the new quality of service requirements of modern transport networks like the ATM;
- Telecommunications and information technologies are converging in such a way that the whole production chain consisting of creation, processing and storing, distribution and displaying of content is merging into one inseparable process – see [3]).

The most important achievement in ODP is the decomposition of the specification into five viewpoints. This is in itself a major strategic achievement because it divides the specification into five separate and non-interfering parts called viewpoints. Each viewpoint must be specified for every system in order to make the specification complete. The viewpoints are:

- The *enterprise viewpoint* specifying the overall requirements of the system;
- The *information viewpoint* specifying the information content in the system and how individual information elements are managed;
- The *computational viewpoint* specifying how the system behaves in order to complete the tasks allocated to it without taking into account how this process is distributed over several machines;
- The *engineering viewpoint* specifying how the distribution takes place and which mechanisms are required in order to do this;
- The *technology viewpoint* specifying how the system is actually built in terms of machine configurations, operating systems, programming languages and systems for operation and management of the final system.

All viewpoints except the enterprise viewpoint are well described in terms of concrete methodologies and languages. The enterprise viewpoint is rather vague because enterprises are so different that it is hard to find a common denominator for them. However, it is possible to produce a rather comprehensive enterprise model for telecommunication systems.

2 General Enterprise Requirements

For ODP systems a specific formal language based on object orientation is developed for each viewpoint. This is easy to do when defining information structures and when describing how computational processes are co-operating in order to perform a common task. ISO has attempted the same for the enterprise viewpoint but, in my opinion, this attempt has not been successful, the obvious reason being that it is hardly possible to give a single definition of what an enterprise is.

The main issues at the enterprise viewpoint are to identify:

- The purpose of the system to be modelled, designed and implemented;
- How this system is located administratively, technically or otherwise within the enterprise;
- How it (or the enterprise) interacts with the environment;
- Policies that the system should support or satisfy, where the policies may be defined by the enterprise itself or forced upon it by the environment (for example a regulation policy);
- Requirements concerning performance, costs, reliability, availability, dependability, etc.

When an enterprise decides to implement or purchase a new system, say, a local area network, it is done having a specific purpose in mind. Unspecified or wrongly formulated or focused purposes are common sources for bad system designs and wasted investments. Hence, a good enterprise specification is the same as having a good and well formulated strategy for the system. A properly defined purpose is the best guiding principle for the modelling, the design and the implementation that follows. It is, for example, common in telecommunications that customers ask for a specific technology, say, an

[1] *ISO = International Standardisation Organization; IEC = International Electrotechnical Committee; ITU = International Telecommunication Union.*

[2] *CORBA = Common Object Request Broker Architecture.*

[3] *TINA = Telecommunications Information Networking Architecture.*

expensive fixed network ISDN implementation, rather than a solution, for example, anything that can provide, as soon as possible and at the lowest possible price, telephony and data transmission at a rate at or above 9.6 kbit/s. This requirement may be solved in several ways, for example, providing ISDN, traditional telephony with modem, mobile telephony or ATM depending on which technology is available. Defined in this way, all these solutions are acceptable for the user. The solution may even be enhanced or replaced as technology is developing, for example, replacing an early telephone solution with ATM when the technology becomes available.

Take GSM as another example. The defined purpose of that system was to provide *Europe with a land mobile system which allowed users to roam all over Europe, maintaining all subscribed services independently of point-of-access, and having only one point-of-subscription and one point-of-billing*. The definition of the purpose thus contains information on the geographic area where the system should be available (Europe), and conditions that have to be met in the geographic area irrespective of network operator, service provider, customs barriers or legislative constraints in each country involved. Among the first items to be solved in order to meet the purpose was to sort out issues like free passage of mobile stations across national borders, in particular equipment containing encryption devices. If these problems were not solved, there would be no reason for designing the system. This is an example of a strategy which goes far beyond the system design itself.

Already in the planning phase, i.e. at enterprise viewpoint, it is often important to state how the system should be administratively embedded in the enterprise, i.e. defining responsibilities, who has access to the system, who can do what to the system, etc. This may for instance define the security levels the system has to support. Normally, a new system is embedded in an existing environment within the enterprise, requiring guidelines concerning how it should interact with this environment. If this point is not addressed early and at a high level, it is likely that the installation of the new system will require expensive adaptation and cause delays. The lack of a clear view on embedding problems seems to be a common cause that so many systems only partially fulfil their intended purpose.

It is in my opinion an enigma how the year 2000 problem was overlooked or ignored by both those who needed the systems, those who specified them, and those who made them. This is the best example we have of the consequences of a complete unconsciousness concerning the enterprise view of the systems to be purchased. And the problem is universal. This illustrates how little concern enterprises are putting in a proper strategy for the systems they make their entire enterprise depend upon.

Every enterprise has an environment and every system within that enterprise has to interact, directly or indirectly, with that environment. These interactions need to be identified and modelled into the system design and implementation. The more indirect these interaction are, the more difficult it is to identify their impact on the design. Similarly, every new system introduced in an enterprise will impact future evolution of the technological infrastructure of the enterprise. Therefore, modelling should allow for open ended design, i.e. design that can easily be extended or fitted into a new environment. The modelling principles of the TINA Consortium are based on this type of approach because telecommunications is one of the fastest changing technologies today. The openness of telecommunication systems is often manifested by a web-like interconnect structure of applications among a large number of individual and very dissimilar enterprises. Closely attached to this web are also the telecommunications and computer infrastructures of the users.

Again we may use GSM as an example in order to illustrate the dependencies between systems and how they may be handled. The embedding requirements were 1) that the GSM system should fit seamlessly into the existing fixed network, and 2) that no redesign of the latter would be necessary in order for them to interwork properly. Both requirements were achieved. However, some of the features of GSM were then designed in a way which was not optimal neither for that system alone nor for the fixed network. The latter has later adopted some of the mobility features of GSM in order to support services like universal mobility. In this case there was an enterprise requirement but it turned out to be suboptimal in the long run. This just illustrates how difficult this type of strategic specification is, and how easy it is to end up with bad solutions.

Every enterprise has a set of policies of its own and has, generally, to meet policies and regulations set by external bodies. Examples of such policies can be related to information security, openness, protection of personal integrity, protection of third parties, rules for co-operation with other enterprises and business conditions. The enterprise viewpoint is the right place to identify which policies and regulations, internal and external, will have impact on the design. At all stages of modelling, design and implementation, it should be verified that policies are not violated. The sector specific regulation of telecommunications is a good example of the impact of the environment. It also shows how difficult it is to understand this interaction for complex systems like telecommunications (see [3] for more details).

The final point in the list above concerns technical issues like performance, reliability, dependability and so on. Any such requirement may have a dramatic impact on implementation choices and cost since they may require duplicated groups of equipment and expensive, high quality parts and components in order to meet reliability specifications, and much spare capacity of computers in order to meet performance requirements for stochastically varying loads. This is particularly true for telecommunications equipment where the impact on the society of failures may be tremendous [3]. However, they have usually little impact on modelling and design.

Much of the success of system development lies in formulating clear and relevant enterprise requirements, and stick to them all the way down to implementation.

The remainder of this essay will contain some general enterprise models for telecommunications. These models are also inherent in every system using telecommunications. From an enterprise point of view the enterprise model for a distributed application using telecommunications for connectivity may be visualised as in Figure 1.

Every distributed system needs the support of telecommunications. This establishes some primary concerns. The enterprise model of the application must interact with the environment of the enterprise itself and with the telecommunications system supporting the distribution of the application and its interactions

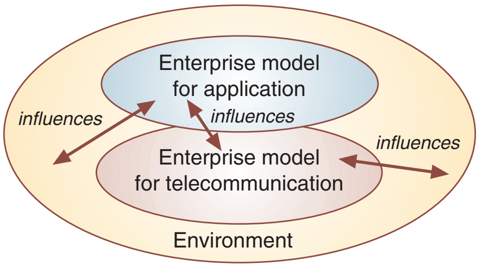


Figure 1 Relationship between enterprise models

with the environment. Some of the interactions with the environment may be direct or via the telecommunications system. When specifying a model it is then essential to define three sets of interactions and their possible interdependence. The latter is probably the most difficult task of the design and probably often overlooked. In particular, it may be difficult to see the secondary relationship imposed by the telecommunications system on the environment and the impact of any limitations in the telecommunications system itself and how such limitations may be bypassed using other means. Designers of applications are often taking telecommunications for granted and do not incorporate it as an integral part of the overall design. For example, when devising a system for electronic commerce, different forms of telecommunications are needed in order to meet the overall goal: web-like systems for browsing of catalogues, transaction systems for payment, security systems for avoiding fraud and supporting non-repudiation, transport channels for delivering electronic goods, etc. This is not a business area where telecommunications is at focus (though many actors in the field act as if that is the case) but where the main driving forces are rather to do more business in the primary area (trade), to be more competitive and to reduce costs and increase profits.

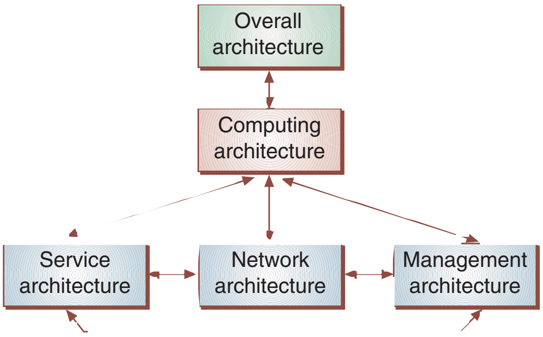


Figure 2 Hierarchy of architectures

Telecommunications will be a support system for many different applications. It will partly be general, that is, applicable for many different applications, or specifically designed for one application. This is what makes telecommunications so difficult and requires that separate enterprise models are developed for telecommunications which may be integrated with enterprise models of very different applications. The skill to do this may become a core competency in telecommunications companies in the future. Few are clever at doing this today.

This is also a challenge for the telecommunications system designer. The system must contain attributes which are general in the sense that they may be used by many enterprises and, at the same time, be adaptable to any specific enterprise and its environment. The enterprise models which are presented below for telecommunications are general and derived with these aspects in mind.

3 Telecommunications Architecture

The TINA Consortium has defined an architecture for telecommunications which separates the system into five interrelated architectural entities as shown in Figure 2.

- *The overall architecture* contains all principles applicable to all designs. This contains all methodology and principles of the ODP viewpoints, including the enterprise viewpoint. As explained above, the same overall architecture can be used for both telecommunications and the application requiring it or, to put it in other words, applications should be specified using similar tools and methods as for telecommunications in order to increase the benefits of the specification and to identify the influences between the enterprises involved.
- *The computing architecture* is a subset of the overall architecture containing all principles for modelling, designing and building the distributed system. This architecture is independent of whether the application is a telecommunications application or a general distributed application. This corresponds to the specifications of the information, computation and engineering viewpoints. This specification is beyond the scope of this essay. The interested reader is referred to [4].
- *The service architecture* uses the methodology of the computing architecture for defining and implementing telecommunication services. This architecture contains the concepts of telecommunications. Here we have to bear in mind that information technology and telecommunications are converging technologies which means that the distinction between them may disappear in the future. This applies in particular to the software used in such systems. However, the service architecture also describes such items as quality of service (bit or word error rate, response time, transfer rate, information loss probability, synchronisation), fault recovery, pricing principles and service interaction handling. These are elements which are absent from most specifications of information systems but are of utmost importance in properly designed telecommunications systems.
- *The network architecture* defines the computational issues of the infrastructure, that is the access network, the transport network and the platforms (exchanges, servers, databases) responsible for routing, processing and service handling. One important point here is that the architecture distinguishes clearly between services and networks. This has two important consequences. First, services and applications can be defined independently of which network they are run on. Second, the fast evolving services and the application technology become independent of the more static network technology. The difference in evolutionary speed of the two will probably become more pronounced because of the low entrance threshold for services and the high investment requirements for networks [3].
- *The management architecture* consists of applications required for managing, monitoring and maintaining the other elements of the system. The architecture is based on the telecommunications management network (TMN) defined by ITU and ISO. This means that TINA specifications use different methods for defining services and management.

As indicated in the figure, several relationships exist between the architectures. It is important to note that all architectures must be in compliance with the computational architecture. This means simply that the modelling principles should be used consistently when defining services, transport and management.

When defining a service or an application, an abstraction of the network will normally be needed in order to show how the service or application interacts with the network resources. This abstraction is given by the network architecture. Note that this abstraction may also hide which elements are realised as hardware or software or which technology is used. This is what enables the separation between services/applications and networks.

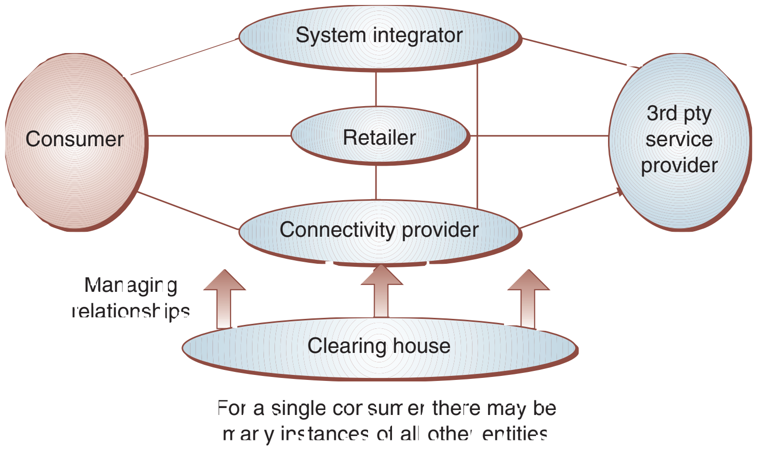


Figure 3 The information service supermarket

4 Business Architecture

An important part of the enterprise viewpoint is to understand the business relationships that are required. This is particularly important for telecommunications since the international network, owned by a large number of companies, offers full connectivity among all its customers and thus requires co-operation between all these companies. In addition, telecommunications is now changing character in that new markets, services and network technologies are evolving and that several new types of companies offering different aspects of telecommunications are entering the market [3].

A business architecture for telecommunications can be built in many ways. In Telenor we have developed a business model consisting of ten atomic roles plus the content provider [3]. The roles are atomic in the sense that it is meaningless to decompose them further and that any company in telecommunications can be present in a single role or be composed of any set of such roles. The TINA Consortium has proposed a similar but simpler model, the *information service supermarket model* [5]. We use this model here because it is well suited for our purpose which is to identify a few primary domains into which the system may be subdivided. The model of Telenor is better suited when designing business models and competition scenarios. The model is shown in Figure 3.

The model consists of the following main elements. The customer purchases services from a retailer reselling connectivity and traffic capacity obtained from connectivity providers and combining them with own or third party services such as customer care, provision of content, storage of content, design of content (web pages), and support of transactions. The structure may become so complex that system integrators are required to keep the structure together. Note that this structure will be able to offer any service from plain telephony to the most advanced multimedia services based on virtual reality delivered jointly by a number of operators. The reasons are, of course, 1) that retailers and system integrators may put together any set of business interests, and 2) that a digital network may pass any type of information provided there is enough bandwidth. The latter may mean that all services and applications may not be available to everybody everywhere. Note that there may be more than one entity of the same type (retailer, system integrator, third

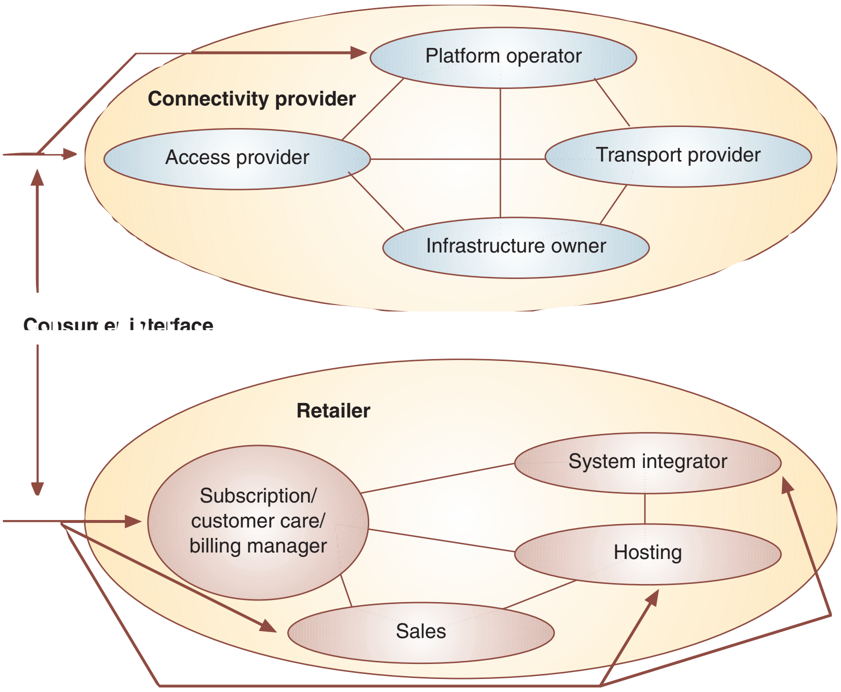


Figure 4 Decomposition of roles

party service provider, connectivity providers and clearinghouse) offering services to one customer.

In a structure like this it may also be possible to combine telecommunications services with other businesses, for example combine electronic commerce with physical delivery of goods and advanced transport logistics where, for example, Rema 1000 sells the goods electronically and Pizza Hut delivers them at the door of the purchaser, and there may be a third company computing the delivery routes (travelling salesman algorithm or the like) and thus reducing the cost of delivery.

The purpose of the clearinghouse may be to establish trust among the parties, act as a *notary public* for proving that transactions have taken place (non-repudiation), offer anonymity and erasure of electronic traces, and perform security functions such as integrity, certification and encryption. The clearinghouse may also offer support for complex transactions.

Figure 4 shows another way of looking at the business system. Here the retailer and the connectivity provider have been divided into several roles. The retailer roles may be sales, hosting (for example offering Internet servers and web hotel), management of customer relations (subscription, billing, customer care) and system integration. The latter represents the complex role of integrating services, networks and content from a number of independent operators. The connectivity provider may be a collection of access providers offering access on copper, fibre or radio; transport providers owning cables, fibres, satellites or multiplexing equipment; platform operators running the machine infrastructure consisting of servers, exchanges and computer nodes; and infrastructure owners owning buildings, ducts, masts and other basic infrastructure required by the others for building the system. This also shows how we may easily superimpose the model in [3] on top of the TINA model. The customer may have business interfaces to these extended structures at none, one or several points. However, physical interfaces may exist independently of business interfaces: a customer may have only the one business interface with a sales organisation. The physical access interface with an access provider is an indirect part of this business relationship.

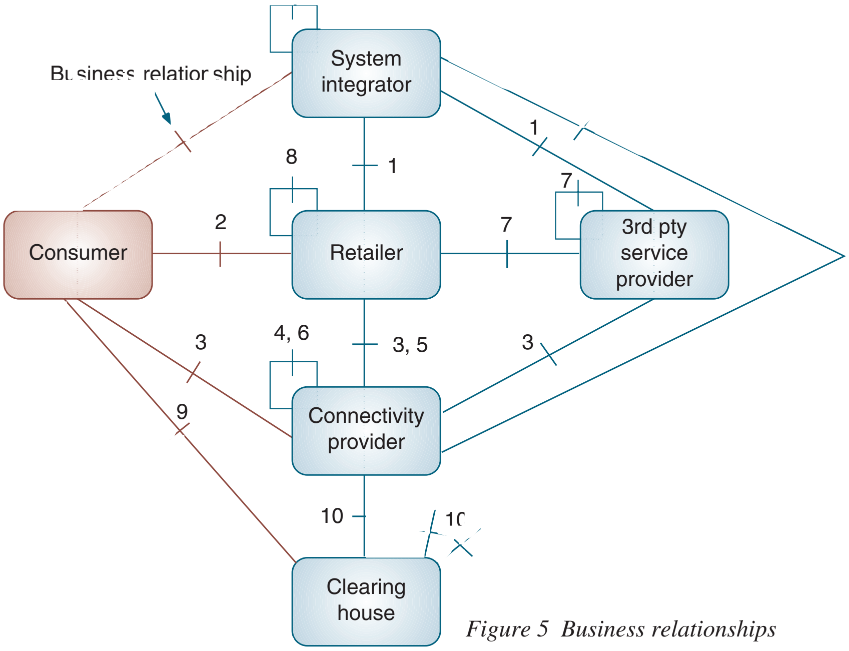


Figure 5 Business relationships

The model is becoming even more complex realising that the consumer may be different things in different contexts: a single person or a company, a machine, a software entity, or a combination of these. The telecommunications service may be a primary service, that is, it is the final product. Person-to-person telephony is an example. The service may also be secondary, that is, delivered as part of another product. In electronic commerce telecommunications is only one (but necessary) component of the offered product and as such a secondary component in the offer. Telecommunications may also be a ternary product in the sense that it is used to produce a product but is not a visible part of the product. Telecommunications used to enable payment transactions in electronic commerce may be regarded as a ternary product since in principle, it can be exchanged with any payment scheme without altering the product as seen by the consumer.

As is seen in Figure 5, the model requires several types of business relationships between the various parties. The figure indicates that as many as ten different business relationships may exist, each of which characterised by a set of attributes and rules. Telecommunications companies therefore see a very complex environment. This figure also indicates that there may be more than one role of each type in the system and that interaction with different parties may be indirect. The customer may, for example, receive a service produced jointly by one third party service provider and several connectivity providers but offered to the customer through an interface of type 1 by a system integrator. The customer will in this case only see the system integrator and not the parties actually producing the service.

The relationship between two business roles may exist over a long period of time (for example, so long as a subscription exists) or only for a single call or part of a call.

The business relationship may be defined in terms of attributes. Take the relationship between the customer and the retailer (marked 2 in Figure 5) as example. The attributes may be:

- Set of services and applications;
- Billing conditions and prices;
- Conditions related to dialogue initiation;
- Set of business domains required;

- Binding requirements;
- Quality of service management.

In an object oriented model these may be attributes of the relationship or the objects (or roles) it interconnects. This item will be considered in [6].

From a strategic point of view it is important that all these relationships are specified for a given configuration. It is particularly important to verify that relationships and attributes are not forgotten in this process.

5 Domains and Domain Interactions

One of the most important aspects of distributed systems is the concept of domain. Every public or private telecommunications system is open, meaning that it interacts with other systems in order to connect customers and provide joint services to them. In this respect the system is divided into domains indicating ownership. Such domains are called *administrative domains*. A business role as described in Chapter 4 may be an administrative domain, or an administrative domain may consist of several business roles if they are owned by the same company. In other cases two systems built to different technologies are required to interwork; that is, to provide joint services to their customers. In such cases the systems belong to different *technological domains*. These are independent of ownership.

The domains need to interact in order to perform their joint task. The interaction need not be symmetric. The configuration of two interacting domains is shown in Figure 6. The possible asymmetry is indicated by assigning different sets of rules (12 and 21) for the two directions.

The interaction between domains takes place via *interceptors*. The interceptors are abstractions of the interfaces between the domains and may act as:

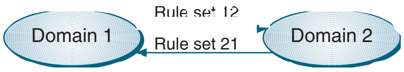


Figure 6 Domain interactions

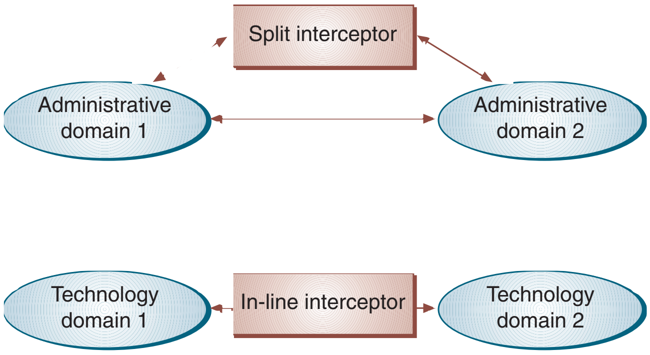


Figure 7 Interceptors

- *Agents* acting on behalf of another domain or entity of that domain;
- *Gateways* performing translations or conversions between domains;
- *Monitors* surveying domains or elements of the domains.

Another way to look at interceptors is in terms of which type of domain they interconnect. In this case two types of interceptors may be defined (see Figure 7):

- *Split interceptors* interconnect administrative domains. The interceptor is responsible for ensuring that security and other policies are met when two domains are connected together. As the name suggests, the interceptor does not take part in the information exchange. It is only needed when security and policy information are exchanged between the domains. It may, however, perform continuous monitoring functions, for example in

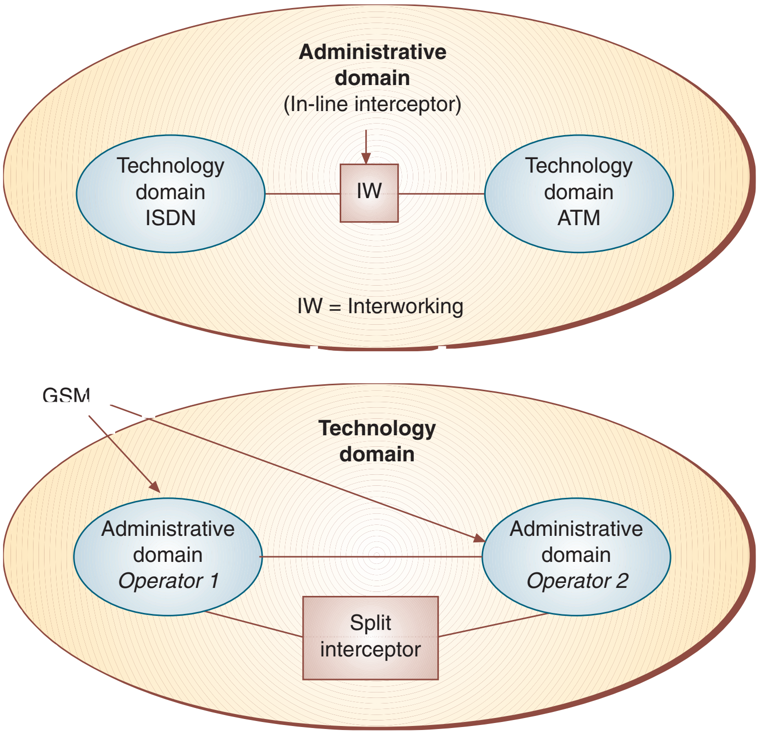


Figure 8 Examples

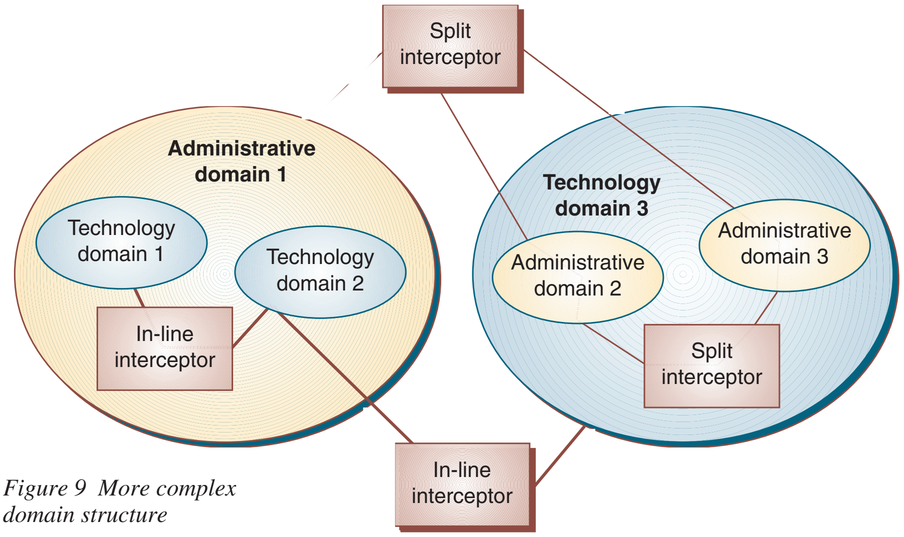


Figure 9 More complex domain structure

order to detect policy violations. The split interceptor may be of type agent or monitor.

- *In-line interceptors* interconnect technology domains for interworking and conversion. The in-line interceptor may be of type gateway.

Figure 8 shows two examples. In the first example two technology domains of the same administrative domain are interconnected. One domain is built with ISDN technology and the other with ATM technology. The interceptor will in this case be of type in-line interceptor and perform interworking of the two technologies. In the other example there is only one technology domain (GSM) but two operators; i.e. two administrative domains are interconnected. A split interceptor is needed where the interceptor essentially takes care of the security between the system and the mobile station, location updating and routing of calls to the mobile station.

Domains may consist of domains as shown in Figure 9. In the example administrative domain 1 consists of two technology domains interacting via an in-line interceptor. Technology domain 3 consists of two administrative domains interacting with technology domain 2 via an in-line interceptor. Between the administrative domains there are split interceptors. It is easy to envisage much more complex situations than this.

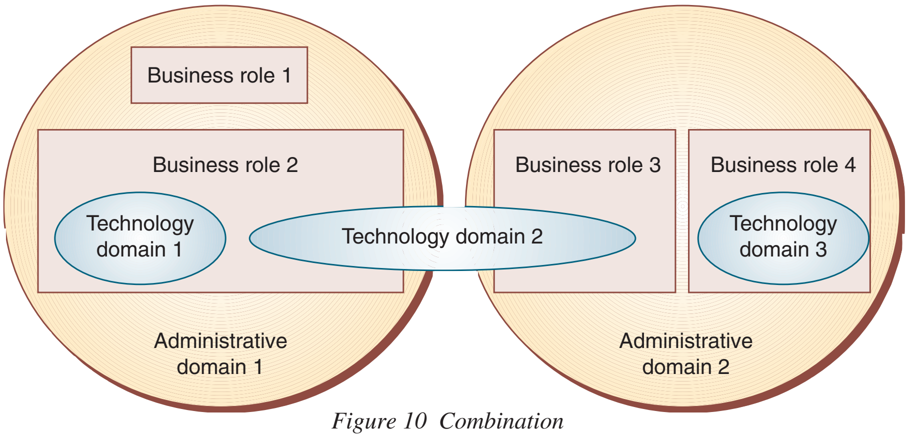


Figure 10 Combination of business model and domain model

The domain concept may be used in the business model. An administrative domain may span several business roles, or several business roles (of the same type or of different types) may belong to the same technology domain. This is illustrated in Figure 10 showing two administrative domains each including two technology domains. One of the technology domains is common for the two administrations. Each administrative domain consists of several business roles. The figure also shows that there may be business roles not containing a technology domain. Most retailer roles may be of this type. Similarly, one business role may offer several technologies, for example a connectivity provider having both ISDN and ATM networks. The example does not show the interceptors. However, there will be a split interceptor between the two administrative domains and in-line interceptors between technology domains of different types (if the technologies are interconnected).

A final point on domains concerns the external visibility of the roles of the business model. A consumer will, for instance, only see the role with which there exists a business relationship and a ‘federated’ role which may be called the telecommunications system with which there may exist a technical relationship (access, connectivity). The federated role and the relationship with it will then have only the attributes which are important for the technical interconnection of the consumer and the telecommunications system. It is unimportant whether the federated role consists of a single entity or is composed of several connectivity providers and third party service providers. In general, it is possible to federate roles in this way in order to hide complexity. For the customer it is unimportant how a service is composed or who provides individual parts of it. Moreover, various sets of different roles may produce the same service. The result is the simple domain structure shown in Figure 11 where the consumer and retailer roles are kept in order to show the subscription relationship but all the other roles have been federated into the telecommunications domain.

Another model that will come out handy in other contexts, for example when discussing middleware and transparency for mobility [7], is shown in Figure 12. Here the retailer domain and the telecommunications domain have been federated and the new terminal domain has been intro-

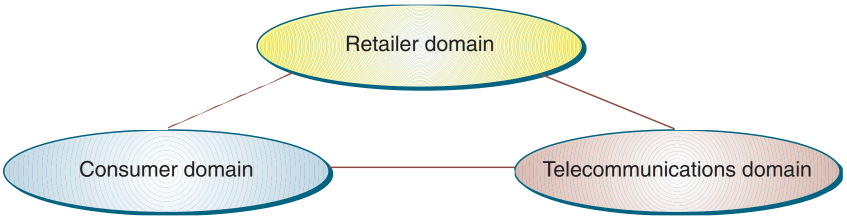


Figure 11 Federated domains

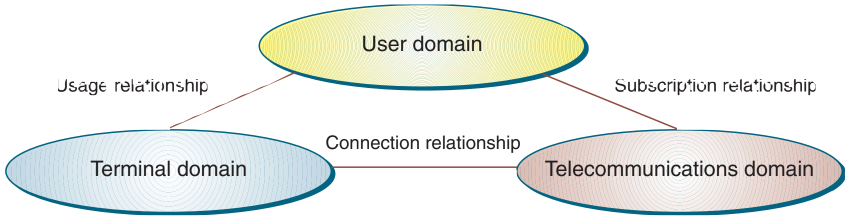


Figure 12 Other federation

duced. The consumer domain has been redefined as user domain. The terminal domain will then contain all user functions and support the technical interface with all other roles. The relationships between the three domains consist of a subscription relationship between the user domain and the telecommunications system domain; a connection relationship between the terminal domain and the telecommunications system domain; and a usage relationship between the user domain and the terminal domain. The model of Figure 12 is thus not strictly a role model but rather a technical one bringing forward three important elements:

- The terminal and the network may belong to different administrative and/or technological domains. In a general model we may assume that the three domains of the figure are different both from an administrative and technological point of view.
- A user is normally not a technical device but needs a terminal on which to run the local applications of the service. Note that a user may not only be a person but also a hardware device or a software entity. However, business information such as user service profile and service restrictions may be required (supported by the subscription relationship between the user domain and the telecommunications domain).
- The terminal is connected to the network. In some realisations it may be regarded as part of the telecommunications domain; in other applications it may be part of the consumer domain. In some applications such as mobility systems, it is advantageous to regard it as a separate domain [7].

In the federated model there may be split interceptors between two and two domains since they are all administratively separate. Between the terminal domain and the telecommunications domain there may also be an in-line interceptor, for example a modem between a PC and the telephone network.

6 Sessions and Applications

In this chapter the domain model of Figure 12 will be used. The telecommunications system offers services which together with other processing make up an application. This is the same as saying that telecommunications should not be, as it often has been, regarded as a stand alone product but as a component of an application also involving other resources. Applications like electronic commerce, distance learning and electronic banking are examples of applications where most of the application is related to things other than telecommunications.

The activity between a user, or rather the supporting terminal domain, and the telecommunications domain is called a *user session*. Activities between two subentities of the telecommunications domain or between telecommunications domains are similarly called *provider sessions*. The overall activity is simply called a *session.* Figure 13 shows a session consisting of user sessions and provider sessions. Using this notion, communications between three parties can be illustrated as in Figure 14. The three interworking telecommunications domains are connected by one provider session in the model.

There are several important points related to sessions. A session must be established before any communications can take place. The establishment of the service may be regarded as a separate subsession called an access session. Sessions are independent of one another. The set of sessions of a telecommunication service or application may change with time; that is, the number of participants and the type and number of resources allocated to the process may vary. Again an organising type of session called communication session may be introduced in order to provide an abstract view of the connection or interdependence between different sessions; that is, providing a connection graph for applications like the one shown in Figure 14.

Instead of talking about end-to-end applications it is better to refer to *incoming applications* and *outgoing applications* where an outgoing application is initiated by the user, while an incoming application is initiated by an entity in the telecommunications domain or by another user, that is, by an outgoing application. A two party call will then consist of one outgoing application and one incoming application interconnected with each other. The incoming application is in this case started by the outgoing application. An example is shown in Figure 15. In a

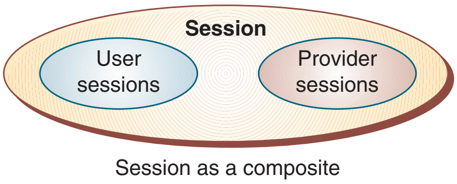


Figure 13 Session and sub-session

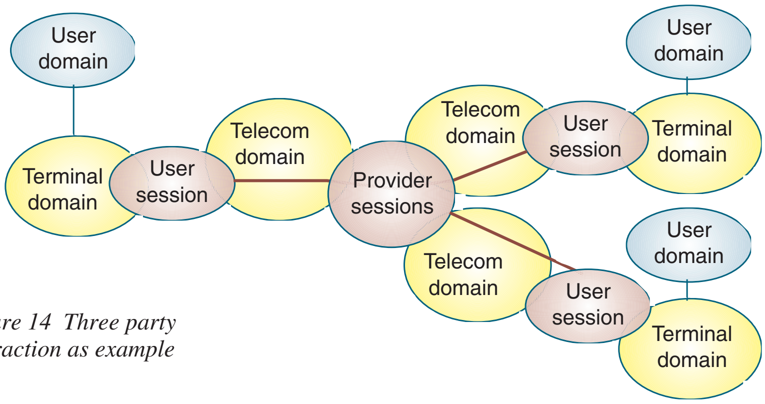


Figure 14 Three party interaction as example

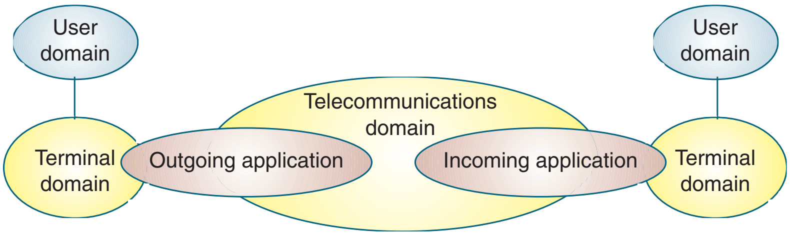


Figure 15 Applications

particular communications event there may be any number of incoming and outgoing applications.

There are several examples where one of the users is the network (or, in order to be formal, the telecommunications domain): the wake up service is an incoming only application; the location updating service in the GSM system is an outgoing only application.

7 Enterprise Specifications and Strategy

The outline given above of the enterprise viewpoint for telecommunications offers a reasonable approach to the modelling of complex systems. By using the proposed models for the structure of the overall architecture, the business relationships, the domain structure and the concepts of session and applications it is possible to reduce the complexity of the system or divide it into manageable pieces. Even though the methods seem simple and almost trivial, this is in fact the most complex part of system design. One very important reason for this is that issues related to enterprise specifications are close to the decision making part of the enterprise. This implies that

- The impact of errors may be great;
- If the enterprise specification is wrong, it does not matter whether the design is correct (the year 2000 problem may be unsolved even though the computer may otherwise work according to specifications);
- Small changes at the enterprise level may require major redesign of the system.

Therefore, it is essential that much effort is put into this part of the specification and not, as has apparently happened in several cases, be ignored completely. ODP offers a top down approach to design where each viewpoint depends on the other in an hierarchical order. As we have seen above, the topmost element of this hierarchy, the enterprise viewpoint, may also be broken up into several areas which may be specified independently.

From a strategic point of view this is unquestionably the most important part of the specification since it will set the requirements for everything else which is said about the system.

References

1 ISO/IEC/ITU. *Open distributed processing : reference model : foundations.* ISO/ICE Rec. 10746-2, ITU-T Rec. X.902, 1996.

2 ISO/ICE/ITU. *Open distributed processing : reference model : architecture.* ISO/IEC Rec. 10746-3, ITU-T Rec. X.903, 1996.

3 Audestad, J A. Telecommunications and complexity. *Telektronikk,* 94, (3/4), 1998, 2–20. (This issue.)

4 Audestad, J A. *Distributed processing in telecommunications.* To be published.

5 Mulder, H (ed.). *TINA business model and reference points.* TINA-C Baseline Documents, 1997.

6 Audestad, J A, Kjæreng, A, Mahieu, L. An object oriented model of telecommunications business. *Telektronikk,* 94, (3/4), 1998, 54–61. (This issue.)

7 Do van Thanh. *Mobility as an open distributed processing transparency.* PhD thesis, University of Oslo, 1997.

Jan A. Audestad (56) is Senior Advisor for the Corporate Management of Telenor. He is also Adjunct Professor of telematics at the Norwegian University of Science and Technology.

e-mail: Jan-Arild.Audestad@s.hk.telenor.no

Agent Nets: a methodology for evaluation of enterprise strategies in information economy

ALEXEI A. GAIVORONSKI

In this paper we address new challenges which are now facing corporate strategic planning in the telecommunication industry. These challenges are the consequence of two interrelated phenomena: the emergence of the information industry as a result of convergence between the telecommunications, computer and content provision industries; and the accompanying deregulation process. Traditional approaches to the evaluation of enterprise strategies fail to produce adequate results because they lack capabilities to represent rapid changes, non-stationarities, uncertainty and complex relations between a multitude of players, which characterise the new industry.

Our answer is to utilise a new methodology for evaluating enterprise strategies, which is based on the emerging theory of distributed multi-agent systems. It takes the best features from more traditional approaches like systems dynamics, and enhances them with the latest advances from modern economic theory, operations research, game theory and discrete event simulation. In order to emphasise the network structure of modern industries it is called *Agent Nets*. It is specifically designed to represent complex systems made up by independent entities called *agents* which transform and exchange information and other resources making independent and co-ordinated decisions on the basis of incomplete information about the state of the whole system and the actions of other agents. In this paper we discuss the applications of Agent Nets for the evaluation of enterprise strategies, present a mathematical description of Agent Nets, describe an Agent Net simulator – MODAGENT – created for simulation of multi-agent systems, and present a case study dealing with the modelling of industrial relations in the information industry.

1 Introduction

In this paper we describe current results of our ongoing research project whose objective is to contribute to the understanding of information economy and develop methodologies and tools which can be used by players in the industry to evaluate their strategies in a time of rapid technological and organisational change. This project originated in connection with standardisation activities of the European Telecommunications Standards Institute, ETSI, and in particular EPIISG, the follow up of Strategical Review Committee number 6 of ETSI. The objective of these bodies was to understand the problems involved in defining, standardising, engineering and implementing the so-called Information Infrastructure with specific focus on the role of the telecommunications industry in this infrastructure. In particular, SRC6 developed the descriptive Enterprise Model of the Information Industry with the aim of evaluating regulatory and industrial scenarios [8]. It became evident that in order to enhance normative activities it is important to support them with adequate modelling tools, which would permit evaluation of enterprise and regulatory strategies in the new industrial environment. The aim of the project described in this paper was to select, adapt and further develop such tools. In the early stages of this project we benefited greatly from the participation and encouragement of Dr. Mario Bonatti from ITALTEL who at that time served as Rapporteur of EPIISG Working Group on the Specification of Enterprise Interfaces.

We started from the premise that one of the most important phenomena of contemporary economy and society is the unfolding of the information industry as a result of convergence of the telecommunications, computer and information content provision industries. This process brings about radically different relations and organisational forms within the industry as well as radically different relations between producers of information, providers of telecommunication infrastructural services, and users of information services. This process is accompanied by rapid organisational and institutional changes, which pose new challenges for both industry regulatory bodies and for individual telecom enterprises, which need approaches and tools for evaluating different regulation and enterprise strategies under various development scenarios.

In order to be useful such planning tools should possess capabilities of reflecting the following phenomena:

- Transition from technology driven industry to market driven industry (at least to a much larger extent than previously);
- New economic phenomena which guide the creation and distribution of information products;
- Transition from a highly homogeneous user population to a highly diversified population of users of information services, according to their social position, habits, mobility, etc.;
- Transition from a tightly controlled monopolistic environment with fairly well defined and few industry roles basically reduced to a network provider and equipment manufacturer to enrich the ever expanding variety of industry roles with blurred boundaries between the provision of structural and infrastructural services and a multitude of loosely regulated and competing players, each assuming different roles;
- Transition from centralised decision making to distributed decision making. This occurs on two related levels: on the level of industrial relations between different economic agents and industry players, and on the level of telecommunication networks which are being transformed from a relatively simple uniform structure to a federation of interacting networks equipped with different levels of intelligence;
- Emergence of a multitude of small and medium sized economic agents in an industry which is accustomed to big players;
- Network nature of industry competition and collaboration patterns;
- An increased importance of dynamics and uncertainty – the state of equilibrium, if ever reached, is determined by the history of the process, and sometimes by decisions made at a relatively early stage.

Traditional economic and organisation theory has difficulties confronting these problems because it mainly addresses the systems in equilibrium and has much less to say about the systems in the state of change and in conditions of uncertainty. We have found that a majority of traditional modelling and decision support tools suffer from the same drawbacks.

For this reason we decided in this project to focus on the emerging theory of distributed multi-agent systems, further developing this theory and its application to the telecommunications industry. This methodology utilises the best features from more traditional approaches like systems dynamics, and enhances them with the recent advances in economic theory, operations research, game theory, and discrete event simulation.

Soon after starting our project we realised that modelling distributed systems of this complexity is very much an open question, although several promising approaches are starting to make the first steps in modern economics, computer science and mathematics. Therefore, we had to organise our research in three interrelated directions:

1. Development of a mathematical methodology for modelling multi-agent systems.
2. Development of an object oriented modelling system based on this methodology.
3. Applications to the modelling of telecommunications systems and information economy.

The rest of the paper is organised as follows: In Section 2 we discuss properties of multi-agent systems and related modelling challenges. In Section 3 we formally describe the notion of Agent Net and discuss some of its properties. Section 4 is dedicated to an informal discussion of Agent Nets. The simulation system MODAGENT developed for simulating Agent Nets with application to modelling of information economy is presented in Section 5. This system is applicable also for modelling of other multi-agent systems found in telecommunications. Section 6 is dedicated to the case study from modelling of relations between network providers and providers of information services over telecommunication networks.

2 Distributed multi-agent systems

The variety of distributed systems can be described as a collection of relatively independent units, here called *agents,* which process information, traffic flows and other resources; exchange information, resources and money; make decentralised decisions which affect the local state of the networks. Taking telecommunications as a reference, we can see that in the case of the global area network such agents are the network nodes which exchange traffic flows and signalling and make routing and congestion control decisions. In the case of local ATM networks such agents are local exchanges with admission control units negotiating resources with users and supervising the user behaviour, while the users themselves can be considered another type of agents.

In the context of the information industry the modelled system is the information economy, and agents are the enterprises and business units involved in creating, producing and distributing information products, network providers, consumers of information products, government and regulatory agencies. Such agents make decisions about the consumption, transformation and exchange of information, products, services and other *resources*, expand their production facilities, and formulate their strategies in order to achieve specific aims. These decisions are taken in an asynchronous and distributed manner. Agents may combine different *roles* within economy, like content provision, brokerage, and delivery of information service.

A more detailed analysis of the information industry as a multi-agent system can be found in [6, 7, 8].

We use the word *agent* in describing these entities in order to emphasise their capability to make independent decisions, communicate with each other, and exchange different commodities. Historically this word is used extensively in economic modelling to describe economic subjects who make independent decisions. More recently the notion of 'intelligent agents' was introduced to describe computer programs operating in a distributed and networked reality like the Internet and acting on behalf of a human making intelligent decisions. In our modelling of systems composed of agents we draw upon both of these notions.

In all of these cases a number of successful approaches for modelling individual agents exists. Modelling systems of agents, however, is a much more difficult task and existing approaches give only partial solutions for relatively simple problems. Still it is a very important problem for the design and engineering of distributed systems. In this paper we consider an approach for modelling multi-agent systems which we call *Agent Nets*.

We start by observing that although multi-agent systems mentioned above are very different by nature, they possess the following common features, which suggest a development of common methodological approach for their modelling:

- *Network structure.* All these systems can be represented as graphs with agents being the vertices of the graph. The edges of the graph represent exchanges and interactions between agents. For example, in the case of telecommunication networks these edges correspond to the links between nodes, while in the case of information economy the edges are supplier-consumer relationships.
- *Transformation and exchange of resources and flows.* In all cases agents possess internal resources, transform input resources and flows, in- and output resources and flows, and exchange traffic flows and resources. In the case of the network nodes such internal resources are represented by processing capacities and buffers, the input flows are received packets and inbound connections, the output flows are outgoing packets and connections. In the case of enterprises internal resources are production capacities which transform input resources into products and services. In the case of end users internal resources are terminals, while input resources are connections, products and services. Agents exchange flows and resources. Examples of such exchanges include exchange of products for money, exchange of signalling between network nodes while establishing a connection, and exchange of information for information. In what follows we consider as *resources* all 'passive' commodities transformed and exchanged by agents, including resources in common sense, but also all kinds of products, services, information and traffic flows.
- *Distributed asynchronous decision-making and control.* All agents make decisions which involve assignment of input resources, treatment of input and output resources, selection of partners for exchange. These decisions are made in an independent and asynchronous way, although they can be coordinated. Examples of such decisions are routing, admission control and congestion control algorithms of telecommunication network nodes; production, development and purchasing policies of enterprises; and selection of service providers by end users. In their selection of decision strategies the agents seek to follow their individual criteria and aims which are generally not coordinated with each other. Therefore, the design of multi-agent systems can-

not be reduced to simple global optimisation criteria, like minimisation of total costs or maximisation of properly defined 'public good'.

- *Dynamics and non-stationarity.* Multiagent systems can exhibit widely different dynamic behaviour due to richness of positive and negative feedback usually present in such systems. They can have many equilibria and switch in catastrophic manner between them. They can exhibit totally chaotic behaviour even in the absence of random disturbances. Design based on the study of steady state behaviour is not adequate because in many cases such systems change constantly. Examples of such non-stationary changes include the rapid development of mobile networks and the explosion of the Internet. The large part of traditional economic theory and modelling is centred on perfect markets in the state of equilibrium. In such systems the operation of market forces smooth out disturbances introduced by uncertainty and leads the system to an ergodic state of equilibrium. In the case of rapid technological change this is no longer the case because relatively small disturbances and decisions with little immediate impact can have self-magnifying properties due to the positive feedback present in the system (technologies with increasing returns) [4]. This leads to non-ergodicity of the system, which requires the modeller to shift the emphasis from stationary to transient behaviour.
- *Uncertainty.* The lack of ergodicity increases the importance of adequate treatment of uncertainty present in the system. There are two levels of uncertainty present in the system: external uncertainty represented by demand patterns, technological change and different kinds of random perturbations; and internal uncertainty due to the fact that each agent makes decisions without the full knowledge of states and actions of other agents. Economic 'particles' (agents, enterprises, countries) do not follow strong laws like the laws in mechanics and physics (for example the law of gravity). They have flexibility to choose different behavioural patterns. Thus, both models of uncertainty and behaviour of economic agents under uncertainty should be included in the system.
- *Complexity.* Traditional economic modelling deals with systems composed from fairly homogeneous agents with similar behaviour patterns. Instead, we needed the capabilities to model a rich variety of relations where the same agents can compete in one field and collaborate in another overlapping field, assume different combinations of industry roles, and possess different knowledge of the state of the whole system. This complexity leads to a multitude of positive and negative feedback in the system, which under different values of system parameters can lead to different equilibria, and even chaotic behaviour. Even without chaos the presence of multiple equilibria leads to catastrophic behaviour; i.e. in certain points the system abruptly switches between different equilibria with an arbitrarily small change of system parameters. An important objective here would be to define 'robust' patterns and regions of stability.
- *Bounded rationality.* Traditional economic theory assumes that economic agents are perfectly rational and their behaviour is governed by maximisation of certain utility functions. Besides these ideals, the 'best case' scenarios, we included in our system some other models of agent behaviour which assume the bounded rationality of agents, i.e. that their decision actions result from the set of heuristics which vary according to the changing of information patterns, environment and goals [3]. Such heuristics may cause instability and they are constantly being evaluated against obtained results, and new heuristics are generated.
- *Imperfect markets.* Market mechanisms are traditionally analysed under the strong assumption that all actions are made simultaneously at known equilibrium prices. The known adaptive (tatonnement) processes, attempting to explain how disequilibrium prices move towards equilibrium, converge only under strong assumptions. One of them is the reversibility of decisions made at disequilibrium prices. This is a critical assumption for various imperfect and artificial markets. The irreversibility of decisions leads to inefficiency of standard tatonnement procedures and thin markets. It calls for approaches when the equilibrium prices can be found without making real decisions at disequilibrium prices. One of them is to create an artificial multi-agent market system, where computers or 'agents' of participants can simulate trading processes and reach equilibrium before making real decisions. The existence of such a market system with agreed rules on how it should function and an anonymous decentralised exchange of necessary information may allow us to reach efficiency for incomplete markets. The design of multi-agent market systems requires special decomposition techniques and the use of (often stochastic) optimisation techniques.

We confront these modelling challenges by utilising and further developing the emerging theory of distributed multiagent systems. In particular, we consider a new modelling and simulation methodology called *Agent Nets*. It is a network like structure with agents being the vertices of oriented graphs whose edges define the exchanges between agents. This structure is associated with appropriate resource space with agents transforming and exchanging resources from this space. Each vertex of the graph is equipped with the set of transformation functions and strategies. The state of the vertex (agent) is defined by the vector of internal, input and output resources. There are dynamic relations which guide the evolution of the state of the agents similar to those considered in the theory of Discrete Event Dynamic Systems DEDS) [9, 12, 17, 21, 28, 30]. Recently developed chapters of this theory which deal with simulation models of asynchronous systems and with interplay between simulation and optimisation are particularly relevant in this context.

We draw upon the experience accumulated recently in modelling and application of network like structures, in particular Petri Nets [1, 27], Neural Nets [19, 16] and Bayesian Nets [2, 25]. Petri Nets are some of the most successful models of Discrete Event Dynamic Systems and are very good for representing asynchronous distributed processes; however, it is difficult to use them to represent nontrivial agents with resource transformation, exchange and control strategies. Neural Nets are useful models for recognition of complex patterns which can be present in multi-agent systems; however, they lack tools for representing distributed decision making. Bayesian Nets are good for processing incomplete information, but again lack a structure for agent representation. Therefore, a new network like model is needed for representing multi-agent systems.

Another source of our insight comes from related work in computational economy and market-oriented programming, which resulted in the creation of several tools for distributed resource allocation in financial and other fields [14, 31, 32, 33]. Still another contribution comes from the evolutionary economics, which studies dynamic interactions between economic agents [11, 23, 26].

In order to cope with uncertainties inherent in multi-agent systems we utilised approaches developed in stochastic programming [5, 10, 13, 15, 20, 22, 24, 29].

3 Agent Nets

An Agent Net is composed of two oriented co-ordinated graphs

$$\{(A,N), (R,P)\} \qquad (1)$$

where

(A,N) – the *agent graph* – is defined by the set of n_a vertices A called *agents* and the set of oriented arcs $N \subseteq A \times A$. Agents possess an additional structure which will be defined separately. Arcs connect those agents which can be involved potentially in *transactions.*

(R,P) – the resource graph – is defined by the set of n_r vertices R called resources and the set of oriented arcs $P \subseteq R \times R$. Quantity and other attributes may be associated with each resource.

For each agent $a_i \in A$ let us define two sets:

$A_i^+ = \{a_j \mid (a_j, a_i) \in N\}$
– the set of – agents from which oriented arcs point to agent ai;

$A_i^- = \{a_j \mid (a_i, a_j) \in N\}$
– the set of – agents to which oriented arcs point from agent ai;

Similarly for each resource $r_i \in R$ we define:

$R_i^+ = \{r_j \mid (r_j, r_i) \in P\}$
– the set of – resources from which oriented arcs point to resource r_i;

$R_i^- = \{r_j \mid (r_i, r_j) \in P\}$
– the set of – resources to which oriented arcs point from resource r_i.

The purpose of the Agent Net is to model the transformation of resources from R by agents from A. Informally, R_i^+ is the set of resources which participate in the transformation which results in resource r_i. In order to transform resources agents make transactions between each other. The set A_i^+ is composed of all agents who possess resources needed in transformations performed by agent a_i. More formal definitions follow.

3.1 Structure of agents

Agent Nets differ from other network like structures, notably Petri Nets, Neural nets and Bayesian Nets, by a more involved node structure, which permits the modelling of different types of agents from real multi-agent systems.

Each agent a_i is a tuple

$$(T_i, I_i, O_i, E_i, S_i, F_i, M_i, D_i) \qquad (2)$$

where

$T_i \subseteq R$ – set of internal resources; these resources are needed to model production capacities and technical capabilities;

$I_i \subseteq R$ – set of input resources; these resources are obtained from agents belonging to A_i^+ in the process of transactions;

$O_i \subseteq R$ – set of output resources; these resources are obtained by transformation from input resources with the help of internal resources and constitute the *offer* of agent a_i to agents from A_i^-;

$E_i \subseteq R$ – set of exchange resources; these resources are exchanged by agent a_i with agents from A_i^+ for input resources.

These resource sets are connected with agent and resource graphs by the set of constraints, in particular

C1. For each $r_k \in I_i$ exists $a_j \in A_i^+$ such that $r_k \in O_j$.

C2. For each $a_j \in A_i^+$ exists $r_k \in O_j$ such that $r_k \in I_i$.

C3. For each $r_k \in O_i$ we have $R_k^+ \subseteq T_i \cup I_i$.

C4. For each $r_j \in T_i \cup I_i$ exists $r_k \in O_i$ such that $r_j \in R_k^+$.

$S_i(t)$ – *state* of the agent a_i which is the vector of quantities of resources from $T_i \cup I_i \cup O_i \cup E_i$. This vector is indexed by members of resource sets of agent a_i and varies with time t. In this way the state $S(t)$ of the whole multi-agent system is composed of the states of individual agents: $S(t) = (S_1(t), ..., S_n(t))$.

$F_i = \{F_i^{jk}(\bullet), j \in O_i, k \in T_i \cup I_i\}$ – the set of transformation functions which define relations of the type

$$y = F_i^{jk}(x)$$

where y is the amount of internal or input resource r_k necessary to obtain amount x of output resource r_j. Similar transformation functions are defined for internal resources and they describe the amount of input resources necessary to obtain the specified amount of internal resource. In this case such functions are used to describe processes of investment and development.

$M_i(t)$ – the information available to agent a_i at time t. Generally, it is some subset of the state $S(t)$ of the multi-agent system obtained with some delay and contaminated by errors.

$D_i(t)$ – the set of strategies of agent a_i at time t. These strategies depend on the state $S_i(t)$ and information $M_i(t)$ and include

- Production strategies which define the amount of output resources to offer;
- Development strategies which define the amount of internal resources to add to existing ones;
- Transaction strategies involving the selection of partner agents to make the offer of output resources to, fixing the amount of exchange resource asked for the output resource, selection of partner agents for obtaining input resources.

3.2 Evolution of Agent Nets

The Agent Net evolves in continuous or discrete time. Its evolution is driven by transactions between agents. Each agent selects its production, development and transaction program according to the set of its strategies and available information. Each transaction involves the exchange of input resources for exchange resources, which change the volumes of these resources in possession of agents. The evolution of Agent Net is described in more detail in [6].

Agent Nets can be studied using graph theoretical methods and methods used in the study of Discrete Event Systems. There are many open research issues which are discussed together with some properties of Agent Nets in [6].

4 Informal discussion of agent models

In this section we informally discuss the concepts described above and show how they can be translated in order to represent economic reality. We again consider the main components of Agent Nets, but give them more informal treatment.

4.1 Resources

These are the elementary entities from which the system is composed. Or, they can be viewed as an alphabet in which the system is described. In our terminology we consider *resources* to be any commodity or entity which is exchanged, satisfied, manufactured or in any other way changed by economic agents relevant to modelling purposes. Thus, besides resources in the economic sense of the word, other examples of resources are money, all kinds of products, services and needs. In our system resources are divided into five types: *exchange resources, input resources, output resources, internal resources* and *final demand.*

Exchange resources. The most common example of such resources is money, but other types can be modelled as well (information, rights, etc.). These are obligatory resources which are always present in the system and whose flow is treated separately from the flow of other resources. This is due to their economic function of exchange and because performance of agents is often measured in exchange balance terms.

Input resources. These are the resources used by agents for creating products and services and satisfaction of needs. For example, in the case of the agent representing an Internet provider, one of the input resources may be the lines which he leases from a telephone company. In the case of the agent representing an Internet user, some of the input resources are a fixed local telephone service and the Internet connection. Input resources are bought by an agent in the market and may be stored.

Output resources. These are the products and services into which agents transform input resources and which are offered to the market. For example, for an Internet provider an output resource is the capacity to provide an Internet connection of given quality, while for a telephone company the output resource is the capacity to provide a telephone connection. From these examples it is clear that the output resources for one agent are the input resources for some other agents. Output resources can constitute the *offers* to the market and they can be stored.

Internal resources. These are resources which are possessed by agents and are necessary for transforming the input resources into output resources. Examples of such resources are qualified manpower or production capacities. For example, for an Internet provider his Internet node would be his internal resource; for an Internet user it would be his personal computer and specialised software; for a telephone company it would be her network. Input resources can be expanded and otherwise developed and should be subjected to maintenance. Money and input resources are needed for both maintenance and development.

Final demand and needs. These are the final resources which drive the economic activity of the system. They are not transformed or exchanged in the system and constitute needs and demands of the end user. What the final resources are very much depend on the purpose of the modelling. Suppose, for example, that we model the penetration of the new telecommunication voice service, like voice over Internet. Then the final resource may be just ‘demand for voice over Internet’ represented by some expert prediction. On the other hand, we might be interested in looking more closely at how this demand is formed according to some hypotheses about behaviour of customers and price and quality structures of competing voice services. In this case the final resource would be ‘the need for voice communication’ measured, for instance, by distribution of time per day for various types of customers. Input resources in this case may be ‘fixed phone connection’, ‘mobile phone connection’, ‘Internet connection’ and ‘other means’.

This resource structure is very flexible and can easily be modified by reassigning resources to different types and aggregation/disaggregation according to modelling needs.

4.2 Agents

Agents transform and exchange resources described previously. We developed generic agent structures which can be specialised in the rich collection of agents by specifying agent parameters for particular purposes. This structure permits us to model a variety of economic players from enterprises to individual users. Such flexibility is important because we needed the capabilities to model agents which combine the multitude of industry *roles.*

Roles. In the rapidly evolving information economy one of the most important issues for a newly emerging company as well as for an established industry leader is which industry roles to assume. Should the established fixed network provider go into providing Internet services or form a strategic alliance with a provider of cable television? Thorough analysis of information industry roles can be found in [8]. After preliminary analysis we understood that all industry roles can be represented in the alphabet of resources described above, i.e. as transformation of a specific set of resources into another set of resources and their exchange. From this resulted the fact that the agents themselves can be represented in terms of this alphabet.

Thus, the *generic agent structure* in our system consists of *resource sets, transformation functions* and *strategies*.

Resource sets. There is a total set of resources for all systems. Each agent is characterised by four subsets of this set, i.e. a set of input resources, a set of output resources, a set of internal resources and a set of exchange resources. Input resources are all resources which are transformed by this particular agent into internal and output resources. For particular agents generated from the general structure some of these sets may be empty. At each moment in time the state of an agent is characterised by available exchange and internal resources and by stocks of input and output resources.

Transformation functions. There are four sets of such functions in the general agent structure: *production functions, development functions, maintenance functions* and *satisfaction functions*. Production functions tell how much of exchange, internal and input resources are needed for producing the given quantity of the output resource. They have the following structure:

$$v_i = \psi(a, v_0) \tag{3}$$

where v_i is the volume of specific input, internal or exchange resource, v_0 is the volume of the output resource and a are production parameters. In the simplest case these functions could be linear; however, we are specifically interested in the case of increased returns and economies of scale. In such a case $\psi(a,\cdot)$ is a concave function which may asymptotically tend to linear with increasing argument. The simplest case of such a function is the following:

$$v_i = a_1 v_0 \frac{1 + a_2 v_0}{1 + a_3 v_0}$$

where the case describes increasing returns when $a_2 > a_3$ and $a_2 < a_3$ corresponds to diminishing returns.

All other types of transformation functions have the same structure (3) as production functions. *Development functions* describe amounts of input resources and exchange resources necessary for expanding production capacities for a given amount. *Maintenance functions* define the amount of exchange and input resources necessary for maintenance of internal resources and stocks of input and output resources. *Satisfaction functions* define the amount of exchange and input resources necessary for satisfying the need of end user.

Strategies. Strategies are actions which agents undertake in order to achieve specific aims. Strategies depend on the amount of exchange and other resources available to an agent and on the information available on the states and strategies of other agents. The general agent structure includes different types of strategies, for example *pricing strategies, development strategies* and *purchasing strategies*. All these strategies may in some cases be derived by solving dynamic optimisation problems. In other cases such strategies can incorporate principles of adaptivity and bounded rationality.

Pricing strategies define the price which an agent offers for its output resources (products and services). In our case study we implemented the principles of bounded rationality as follows. Each agent had a set of several strategies: keep the market price, increase the price or decrease the price based on previous history. At each step an agent could choose from one of such strategies according to probabilities which were updated according to their performance in terms of income and revenue, similar to the theory of learning automata.

Development strategies. If demands exceed production capacities an agent can choose between increasing the price or expanding production capacities. Development strategies govern such expansion taking into account the fact that newly added capacities are becoming operational after some delay.

Purchasing strategies. These strategies are employed by the agents for the selection of offers for required input resources present in the market. The simplest strategy is, of course, to choose the offer with lowest price. We take into account, however, that for real economic agents the price considerations are not necessarily unique and allow customers to migrate between offers with different prices with some dynamics dependent on other attributes of an offer.

At this moment we implemented some basic set of strategies which is in the process of expansion.

Specific agents are generated from this general agent structure by specifying its elements. Here are some types of the agents with which we experimented.

Production agent. This agent puts on the market products and services produced from production capacities using input resources bought on the market, but does not have final needs to satisfy. These agents further differ by their set of strategies.

End user agent. This agent satisfies the final needs by purchasing products and services on the market. This agent is further characterised by the capability to substitute different products to satisfy the same need. For example, the need for voice communication can be satisfied by fixed phone, mobile phone or voice over Internet.

Pure supplier. This agent has only an output resource which supplies for the price derived from the expert estimates. This agent is useful for modelling the supply of important products whose flow we do not want to describe in much detail for modelling purposes. One example is the regulatory commission which distributes frequencies for broadcast transmission.

4.3 Market

The market is the environment in which agents operate. At any time agents who produce output resources put their *offers* on the market. Each offer consists of quantity and price of the specific resource. Agents who are customers for the input resources go to the market and choose between offers. For the case when demand exceeds supply the system has the set of rules which sees to available supply being distributed between customers. Producing agents may then decide to increase the price for the next period and/or to expand production capacities. There is a set of balancing mechanisms which are needed because the unsatisfied demand of one agent may result in a decrease of its offer to the market which in turn may result in diminishing satisfaction of demand of another agent. One possibility is to use the generalisations of Walras tatonnement process [33].

5 System MODAGENT

In this section we describe the architecture of the object oriented modelling system MODAGENT which implements the concept of Agent Nets described in the previous two sections. It is conceived as a tool for modelling complex distributed multi-agent systems discussed in Section 2.

System MODAGENT shares the methodological approach with system INFOGEN developed at the first phase of this project with support of Italtel, and constitutes its further development. In [6] one can find the description of INFOGEN and the first applications to the modelling of information economy.

MODAGENT, however, proved to be more than just a software tool: it constitutes a contribution in its own right to the modelling methodology of complex multi-agent systems.

Importance of object oriented approach. The complexity of the economic system under study is such that it is unrealistic to pursue the aim of creating a modelling system which would embrace all problems of interest. What is possible, however, is to create a generic modelling system whose structure would reflect the structure of the telecommunication economy, and which would be endowed with the capability of filling this structure with particulars of the given problem. It is important that different components of the system can be changed and specified without involving other components. The object oriented approach provides the appropriate methodology for doing just this, having the means to define classes which encapsulate specifics, hiding them from other parts of the system.

It also provides the means to develop specific models from generic instance by utilising the concept of class derivation. This makes it possible, for example, to define the generic structure of the economic agent and derive from it specific agents with relatively little effort. In addition, object oriented approach encourages the developer to think in terms of the general structure of the application field which results in descriptive models giving important feedback to both mathematical modelling and applications. These considerations determined our choice of object oriented paradigm for development of MODAGENT which is implemented in C++.

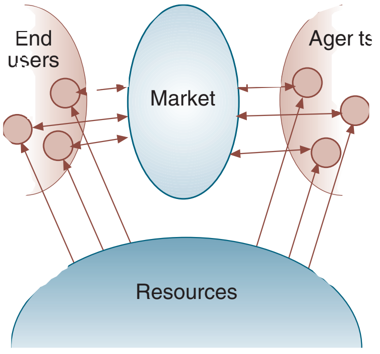


Figure 1 Top level structure of **`Economy`**

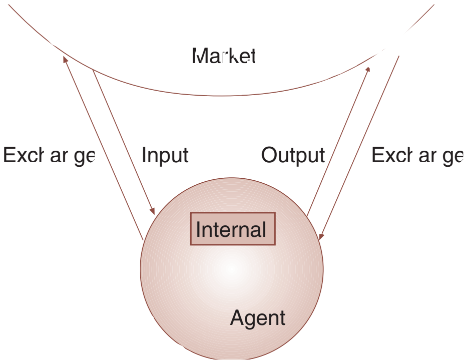


Figure 2 Relations between resource classes and classes **`Market`** *and* **`Agent`**

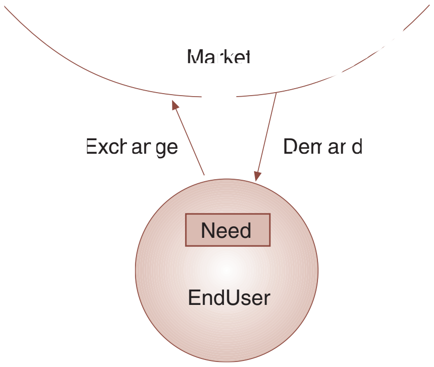


Figure 3 Relations between resource classes and classes **`Market`** *and* **`EndUser`**

5.1 General description of modelling choices in MODAGENT

These will be illustrated in terms of our principal motivation: modelling of information economy. Our approach, however, can be applied to a wide variety of complex distributed systems. We consider economy to be made up by the following four components (see Figure 1):

- *Resources* which are the basic building block of an economy. They are understood in very general terms and embrace all 'passive' entities found in economy: resources in the proper sense, products, services, information and derivatives, demand, consumer needs. Resources can be transformed from one to another, exchanged, consumed, accumulated;
- *Agents* which are the basic 'active' entity of economy. They transform, produce, sell, consume, accumulate, offer, exchange resources, obtain information on economy, formulate and execute strategies in order to fulfil their objectives;
- *End users* who are a special kind of agents, but which we decided to treat separately for modelling purposes. They consume products and services produced by agents in exchange for money and possibly other exchange resources. In doing so they pursue the aim of satisfaction of their needs;
- *Market* facilitates and organises exchange between agents and between agents and the end users.

These main components constitute the top level classes of the object oriented hierarchy of MODAGENT. In what follows we indicate the names of the classes like **`this`**. The object of class **`Economy`** is to organise interaction between objects of classes **`Agent`**, **`EndUser`** and **`Market`** and simulate dynamics of this interaction in discrete time.

Resources can belong to one of five top level classes: **`Input`**, **`Output`**, **`Internal`**, **`Demand`** and **`Need`**. Besides, there are exchange resources which are handled by special class **`Accountant`**. **`Input`** resources are procured in the **`Market`** in exchange for exchange resources, are transformed by an agent in **`Output`** resources using **`Internal`** resources and are offered to the **`Market`** in exchange for exchange resources. **`Need`** resources are satisfied by the end user consuming **`Demand`** resources procured in the **`Market`** in exchange for exchange resources. Thus, resources belonging to the **`Input`** class of one agent may belong to the **`Output`** class of another agent and to the **`Demand`** class of the end user. The relations between different resource classes and classes **`Agent`**, **`EndUser`** and **`Market`** are shown in Figures 2 and 3.

The complex relations which connect agents and end users necessitated a further structuring of **`Agent`** class into high level subclasses. Besides resource classes **`Input`**, **`Output`** and **`Internal`** mentioned above, these classes include:

`Information` handles information available to **`Agent`**;

`Consumer` procures and handles **`Input`** resources;

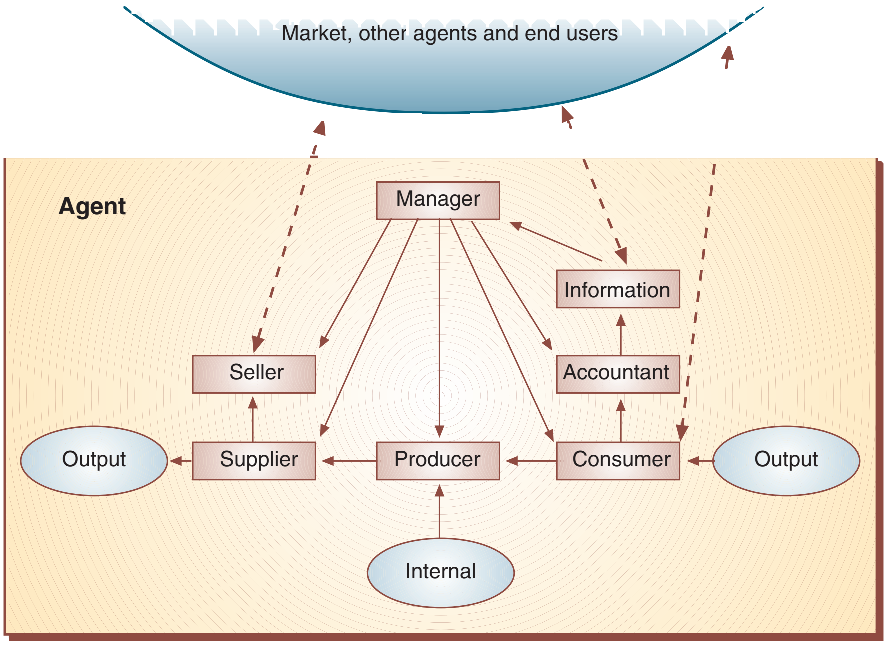


Figure 4 Top level structure of class **`Agent`**

`Producer` transforms **`Input`** resources into **`Output`** and handles expansion of **`Internal`** resources;

`Supplier` handles and dispatches **`Output`** resources produced by **`Agent`**;

`Seller` markets **`Output`** resources;

`Accountant` handles the flow of exchange resources, keeps balance and generally keeps track of **`Agent`** performance;

`Manager` defines policies executed by other members of **`Agent`**.

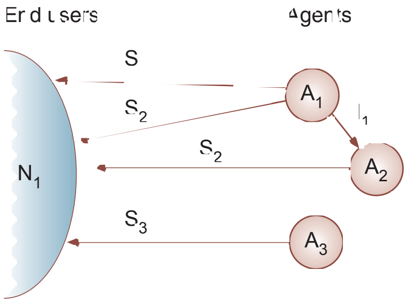


Figure 5 Providers of infrastructural and structural services

The structure of these classes will not be described here. The top level structure of **`Agent`** is shown in Figure 4. Many agents will not possess fully developed components of this structure; in such cases some components will be empty.

Detailed description of the MODAGENT structure can be found in [18]. The next section is dedicated to a case study which refers to dynamics of relations between providers of basic telecommunication services and providers of information services which are built on top of telecommunication networks.

6 Numerical experiments with MODAGENT: dynamics of relations between providers of structural and infrastructural services

This section is dedicated to MODAGENT modelling of industrial relations in the emerging information industry. In particular, we model the combination of structural and infrastructural roles by different industrial players. This combination may result in complex dynamic patterns of relations between industrial agents which may include simultaneous competition and collaboration. These patterns are further affected by different industrial strategies of agents resulting from incomplete information in their possession, different relative strengths and weaknesses. These experiments show that the system in its current state already permits us to perform modelling and qualitative analyses of these phenomena, explore dynamics of agent learning, describe complex relations of competition and co-operation between providers of structural and infrastructural services, highlight roles of regulation, and evaluate various market strategies.

6.1 Model of competition and collaboration between providers of infrastructural and structural services

The economic system in question is made up from the following economic agents (see Figure 5).

1. ***End Users.*** They have need N_1 which we here call 'need for information services' for whose satisfaction they have some fixed renewable budget. This need can be satisfied by information services S_1, S_2 and S_3.

Service S_1 describes traditional information services which require infrastructural services of network operators and are provided by network operators themselves, like obtaining any kind of information by traditional telephony, yellow pages, etc.

Service S_2 refers to a new kind of information services which also require infrastructural services of the network operator, but which can be provided by both network operators and independent companies which lease lines from network operators. Think, for example, about Internet based provision of information services.

Service S_3 refers to information services relatively independent of infrastructural services of network operators, like, for example, radio and TV services.

These services are partially substitutable between each other. End users choose between different services according to the following demand generation model. The attitude of each end user towards any particular service S_i is characterised by the service utility function $f_i(x)$ which is a non-decreasing function varying from 0 to 1 and defines the fraction of need N_1 satisfied by amount x of service S_i. This function is characterised by two parameters:

a_i – the maximal possible fraction of need which can be satisfied by a given service when its amount x tends to infinity, we assume that all a_i sum up to 1.

b_i – the incremental fraction of the need satisfaction by service S_i for small values of x. The end user chooses between services by maximising his need satisfaction within the given budget (perhaps with some error). This permits us to model a non-homogeneous user population with the following service adoption cycle. Suppose that there is some established service and some emerging service. In the early stages of service introduction there will be some fraction of 'early bird' users which will adopt a new service even if it costs substantially more than traditional services, it is enough that the new service has some new appealing feature. In the later stages the main body of users will adopt the new service, but only if it becomes price competitive with the old service. And finally, there will be a fraction of users who will switch to the new service only if it becomes substantially cheaper than the old service.

Normally, there is more than one provider of the same service on the market who offer services with different prices. We assume, however, that there are more attributes to a service than the price; services from different providers may differ in many other respects, like quality, etc. Thus, not all end users select the provider with the cheapest service. However, the price remains an important attribute of a service and market quotas of different producers increase and decrease depending on price. This permits us to differentiate between basic services with standard quality and more expensive services with enhanced quality.

Finally, we take into account one more characteristic of end user behaviour: even if services are identical end users follow with some delay the relative price dynamics of different producers.

2. ***Producers.*** There are three groups of producers in our economy:

2a. *Provider of structural and infrastructural services* A_1. He provides structural information services S_1 and S_2 and infrastructural service I_1 necessary for provision of S_1 and S_2. Alternatively, we refer to this agent as *Network Operator*. We assume that provision of service S_2 besides requiring infrastructural service I_1 also requires internal capability C_1 which needs to be developed. All other resources necessary for provision of services S_1 and S_2 are expressed in terms of money.

2b. *Provider of structural services* A_2. He provides information service S_2 and needs to buy infrastructural service I_1 from A_1. Alternatively, we refer to this provider as 'Information Service Provider'. Thus, he is a competitor to A_1, although in order to compete he needs the service I_1 provided by A_1. Take, for example, as A_2 an Internet provider who leases lines from a network operator to provide information services, and take as A_1 a network operator who is also engaged in providing information services. Think, for example, of AT&T or 'Baby Bells' from one side and America OnLine and CompuServe from the other side.

2c. *Provider of structural services* A_3. He provides information service S_3 and is independent of both A_1 and A_2. Think, for example, of a newspaper publisher or an owner of TV channels. His economic function is to provide a service which can substitute to some extent both S_2 and S_1 in case they become too expensive or inadequate in some other respect. We assume for simplicity that he provides service S_3 for a fixed price.

3. ***The objectives of economic agents.*** We assume that end users maximise the satisfaction of their needs for information services as described above. Providers maximise their profit by choosing among different strategies. These strategies belong to three classes:

- Increase market offer, decrease price and expand provision capacities if necessary;
- Decrease market offer and increase price, maybe moving to a more specialised market niche;
- Keep the price and market offer constant.

Different strategies from the same class differ quantitatively, for example by the amount of price decrease or increase.

Providers choose between these three strategy classes according to information about the market behaviour. If profit increased during the current period then the current strategy is maintained or enforced; otherwise the strategy is likely to change. This is done according to the *learning process* which uses information about market behaviour and is organised similarly to the learning process in the theory of learning automata. That is, each strategy is characterised by probability according to which this strategy is chosen in the next step. These probabilities are updated each period: the probability of strategy which improves some specified performance criterion is increased while probabilities of other strategies are decreased. However, probabilities of all strategies remain positive. This approach permits us to model a rich set of possible behaviours.

We took the current profit as the agent performance measure. The approach just described permits us to reflect the fact that the profit maximisation is the most important, but not unique driving force of agent behaviour. Therefore, we also permit keeping the strategy which led to profit decrease, but with smaller probability compared to the case when this strategy increased the profit.

We considered the cases when provider A_1 varies his offer of both infrastructural service I_1 and structural service S_2 keeping fixed the offer of structural service S_1, while provider A_2 varies the offer of service S_2.

Our objective was to model dynamics of competition and co-operation between providers A_1 and A_2, penetration of service S_2 and dynamics of total market for information services S_1 and S_2 under different scenarios about:

- Market and expansion strategies of both providers;
- Different provision costs for service S_2;
- Different policies of market regulation.

6.2 Description of scenarios

Each scenario is characterised by the following components:

1. Relative production costs for provision of service S_2 by providers A_1 and A_2. Here we consider two alternatives: costs of provider A_1 ('Network Operator') are considerably higher then the costs of

Scenario	Provision costs	Policies	Regulation	Maintenance costs
A	A_1 high, A_2 low	A_1 weak, A_2 strong	No	Low
B	A_1 high, A_2 low	A_1 strong, A_2 strong	No	Low
C	A_1 high, A_2 low	A_1 strong, A_2 weak	No	Low
D	A_1 high, A_2 low	A_1 strong, A_2 weaker than in C	No	Low
E	A_1 high, A_2 low	A_1 strong, A_2 weaker than in D	No	Low
F	A_1 high, A_2 low	A_1 strong, A_2 weak as in E	Bounded price of S_2 for A_1	Low
G	A_1 high, A_2 low	A_1 strong, A_2 weak as in E	Low maximal price for I_1	Low
H	A_1 high, A_2 low	A_1 strong, A_2 weak as in E	High maximal price for I_1	Low
I	A_1 same as A_2	A_1 weak, A_2 strong	No	Low
J	A_1 same as A_2	A_1 strong, A_2 strong	No	Low
K	A_1 same as A_2	A_1 strong, A_2 weak	No	Low
L	A_1 same as A_2	A_1 strong, A_2 weaker than in K, same as E	No	Low
M	A_1 same as A_2	A_1 strong, A_2 weak as in E	High maximal price for I_1	Low
N	A_1 same as A_2	A_1 strong, A_2 weak as in E	Low maximal price for I_1, half as in M	Low
O	A_1 same as A_2	A_1 strong, A_2 weak as in E	Minimal price for S_2	Low
P	A_1 same as A_2	A_1 strong, A_2 strong	Minimal price for S_2, maximal price for I_1	Low
Q	A_1 same as A_2	A_1 strong, A_2 strong	No	High
R	A_1 same as A_2	A_1 strong, A_2 strong	Minimal price for S_2	High
S	A_1 same as A_2	A_1 strong, A_2 strong	Minimal price for S_2, maximal price for I_1	High

Figure 6 Summary of scenarios

provider A_2 ('Information Service Provider') and the case when their costs are roughly equal.

2. Capability of the provider to change the market offer from one period to another. This involves the price increase or decrease, possible capacity expansion and capability to add value to a service in case of the price increase. If the provider possesses the capability (or willingness) to vary his market offer considerably we say that he pursues 'strong' policies, if he is capable only of relatively small variation then we say that he pursues 'weak' policies. Obviously, weakness or strength is relative between providers. The strength of the provider is measured numerically by the maximal possible price changes during a certain period of time.

3. Market regulation which in our case consists in fixing allowable price range for services by outside authority or by agreement between providers. We consider the following cases:

- Absence of market regulation;
- Upper bound is fixed on the price of infrastructural service I_1;
- Lower bound is fixed on the price of structural service S_2;
- Both upper bound on price of I_1 and lower bound on price of S_2 are fixed.

4. Costs for maintaining provision capability of service S_2. We wanted to consider maintenance costs separately from production costs because in some cases service providers tend to overdimension provision capabilities in order to increase competition flexibility. We considered two cases:

- Low maintenance costs;
- High maintenance costs.

The model permits us to study also the cases with different economies of scale for different providers, different end user preferences and different initial market conditions which we do not consider in this paper, however. Each individual scenario can be obtained by choosing one of the scenario components described above. We considered 19 scenarios indexed by letters A–S which are summarised in Figure 6.

For each of these scenarios we simulated the economic system described above starting from the same initial conditions: initial provision capacities for service S_2 and initial prices for service S_2 which were equal for both providers. Results of simulations contain evolution of the following quantities:

1 Profit of both providers;

2 Revenue of both providers;

3 Price for information service S_2 for both providers;

4 Volume of service provision for service S_2 for both providers;

5 Provision capacities for service S_2 for both providers;

6 Revenue breakdown for A_1 \$ between services S_1, S_2 and I_1;

7 Total market value for information services S_1 and S_2 with market shares of providers A_1 and A_2;

8 Price for infrastructural service I_1.

Evolution of these quantities are shown in figures named 'Figure X.Y' where X indicates the number of quantity as given above and varying from 1 to 7; Y varies from A to S indicating scenarios defined above. Due to the lack of space we give here only a part of these figures.

6.3 Analysis of results

Results exhibit many interesting phenomena which have convincing economic interpretations, in particular

- Benefits of market regulation which may be performed either by appropriate regulatory body or by agreement between providers. Such regulated competition permits providers to survive who offer superior services but cannot expand the offer rapidly, in many cases safeguarding the total market value and in many cases being beneficial to all players since it stabilises the market;
- Tendency of Network Operator to concentrate on his core business when his information service provision costs are high and his tendency to compete in information services when his costs are comparable with Information Service Provider;
- 'Strong wins': importance of pursuing aggressive market and expansion policies (for those who can afford them), the provider with strong policies assures himself the largest profit share even with somewhat inferior provision costs;
- High instability of non-regulated market when everybody pursues strong policies, price wars;
- Emergence of qualitatively different dynamical patterns of the system evolution which in some cases may include oscillations in vicinities of different equilibria and in other cases markedly chaotic behaviour.

Let us consider in more detail some of the simulation results for different scenarios described above. For the main part we are here going to analyse the evolution of profit, since in this experiment the profit maximisation was taken to be the ultimate goal of providers.

1. Group of scenarios when Network Operator A_1 has higher provision costs for service S_2 compared to Information Service Provider A_2 (scenarios A–H)

These scenarios cover cases when A_1 does not have sufficient knowhow to provide the competitive service S_2 for different economic and organisational reasons. Figures 1.A, 1.B and 1.C illustrate the point of importance of strong policies. By strong policies we here mean the capability of the provider to rapidly change the price of the offered service and, consequently, meet demand variations. Since, in the case of price decrease, demand may increase considerably this also means the capability of the service provider to make sufficient investment in order to meet increased demand. On the contrary, we are speaking about weak policies when a service provider varies the price and quantity of offered service by relatively small steps.

Figure 1.A shows that in the case when provider A_1 employs weak policies and provider A_2 employs strong policies the profit share of A_2 gradually increases while the profit share of A_1 gradually decreases. Profits of A_1 increase compared to profits of A_2 when both providers use the strong policies (Figure 1.B). The profit share of A_1 increases even more when he uses the strong policies while A_2 uses the weak policies (Figure 1.C). In all cases the system reaches equilibrium state after some time and continues to oscillate around equilibrium state. The equilibrium profit distribution, however, depends dramatically on the relative strength of policies employed by providers: the stronger they get, the larger the share. Analysis of volumes and different revenue items of provider A_1 shows that after the initial attempt to compete with A_2 on the provision of service S_2 he decides to concentrate on his core business, provision of infrastructural service I_1, and gradually pulls out of providing S_2; the faster he does it, the stronger his policies are (Figure 6.B). This is understandable, given that his provision costs are higher than for provider A_2. In this way provider A_2 obtains the major part of the market for information services S_1 and S_2, while provider A_1 retains only a small part of

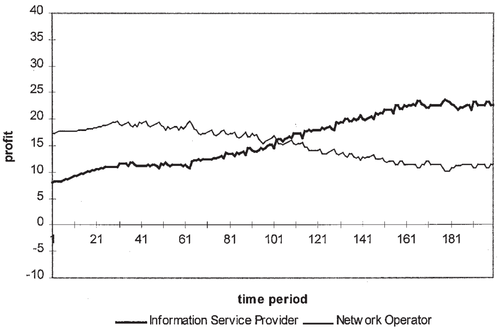


Figure 1.A

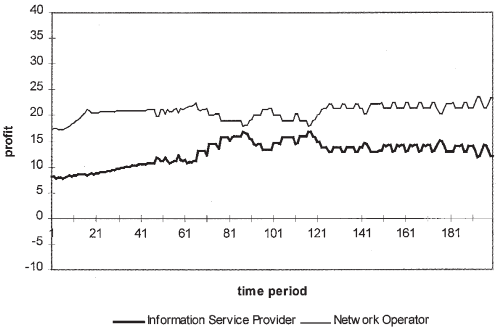


Figure 1.B

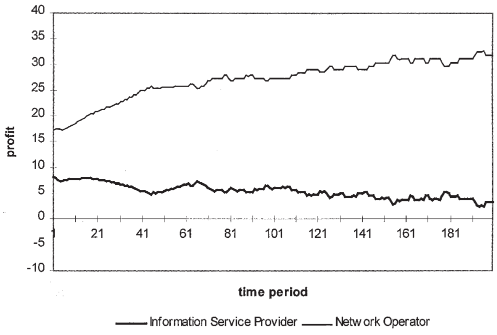


Figure 1.C

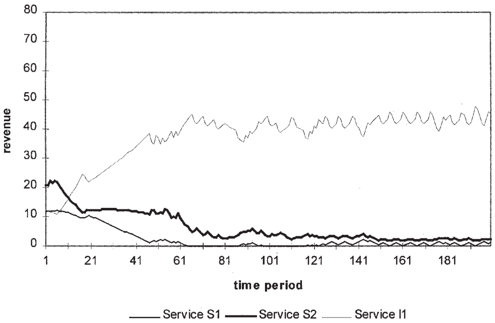


Figure 6.B

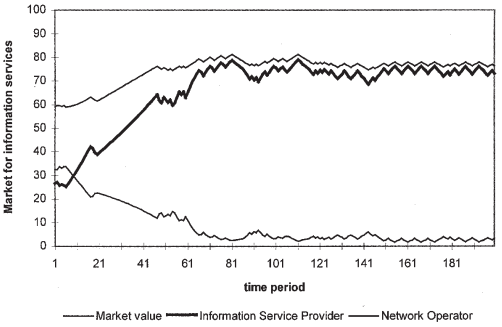


Figure 7.B

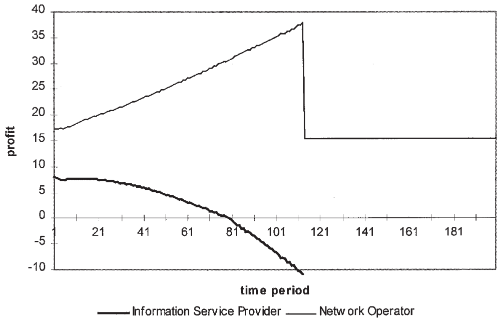


Figure 1.E

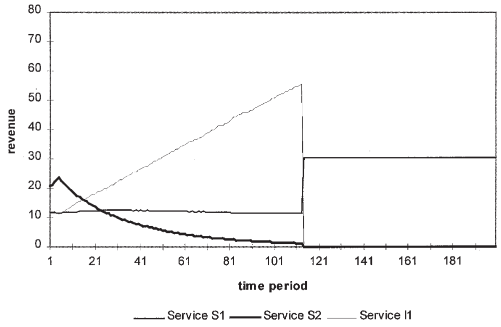


Figure 6.E

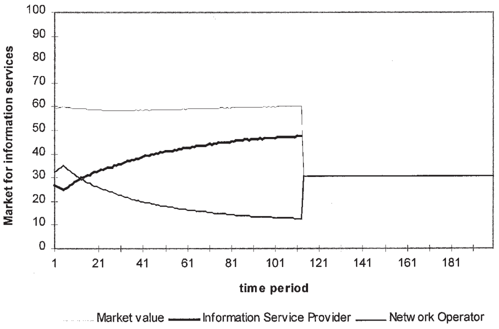


Figure 7.E

the market related to traditional service S_1 (Figure 7.B).

What happens if provider A_2 is only capable of using even weaker policies? In this case, starting from some point, provider A_1 is capable of driving provider A_2 into bankruptcy by increasing the price for service I_1 (Figure 1.E). This has unfortunate consequences even for provider A_1 since he loses his revenue from providing infrastructural service I_1 and is not capable of substituting the lost profit by providing service S_2 himself. He is capable of partially recovering the lost revenue by increasing provision of service S_1 (Figure 6.E), but many users migrate to service S_3 offered by provider A_3. Consequently, both the profit of A_1 and the total market for services S_1 and S_2 decrease substantially (Figure 7.E).

Let us now take a look at the effects of regulation in this case. The first possibility is to force provider A_1 to maintain the capability for provision of service S_2. This somewhat preserves the market for S_2 even in case of the bankruptcy of A_2, but both the profit of A_1 (Figure 1.F) and the total market for information services S_1 and S_2 (Figure 7.F) decrease substantially, although not to the same extent as in the absence of regulation.

Another possibility of regulation is to impose the upper bound on the price of infrastructural service I_1. This permits provider A_2 to survive even in the case of weak policies. This brings

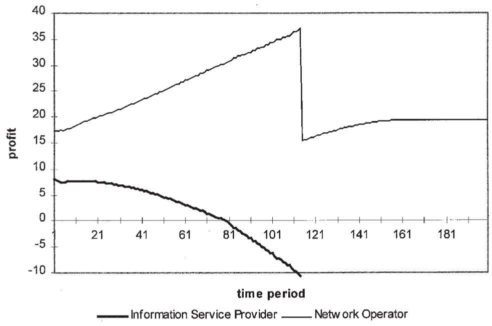


Figure 1.F

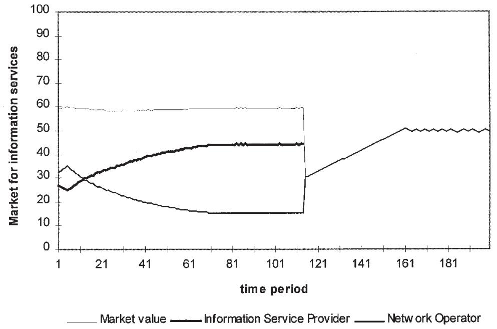


Figure 7.F

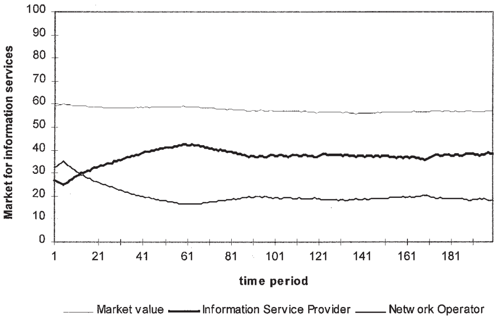


Figure 7.G

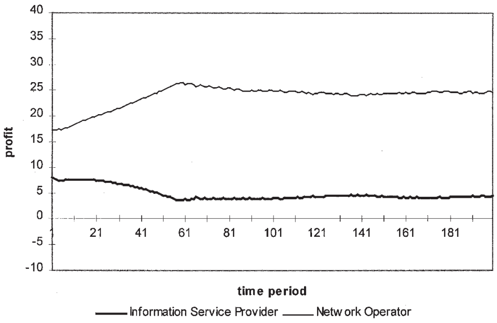


Figure 1.G

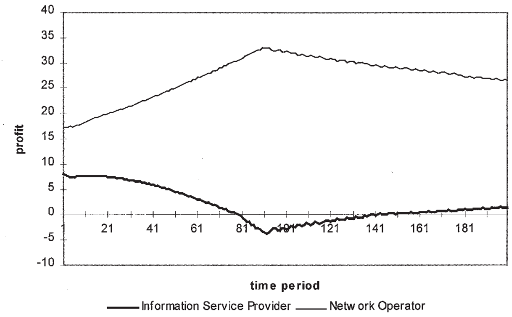


Figure 1.H

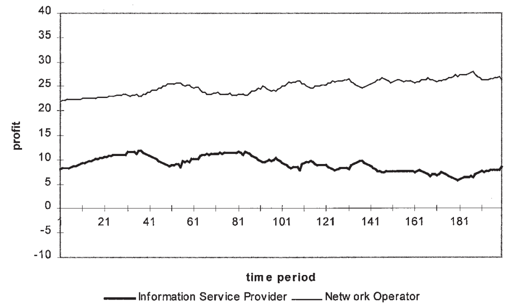


Figure 1.I

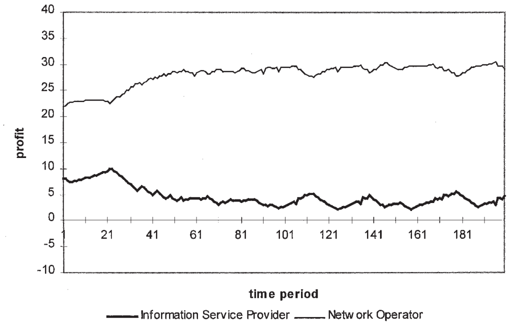


Figure 1.J

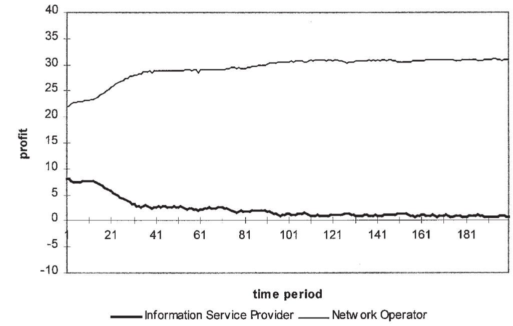


Figure 1.K

benefit also to provider A_1 and increases the total market for information services (Figure 7.G) since the system reaches equilibrium with larger profits for A_1 (Figure 1.G) than in the case of bankruptcy of A_2. Figure 1.H shows a borderline situation where the upper bound for the price of I_1 is higher than in the case depicted in Figure 1.G. Provider A_2 starts to lose money, but finally manages to recover.

Note that this price restraint can be imposed by the regulator body, but it can also be a result of self-restraint of provider A_1.

2. Group of scenarios when Network Operator A_1 has similar provision costs for service S_2 compared to Information Service Provider A_2 (scenarios I–P)

These scenarios cover cases when A_1 is capable of competing with A_2 even on service S_2. Figures 1.I, 1.J and 1.K are similar to Figures 1.A, 1.B and 1.C respectively, and show that in this case profit distribution is shifted substantially in favour of provider A_1, since in this case he gets revenue both from provision of S_2 and from provision of the infrastructural service I_1. Similar to the previous case, the equilibrium profit distribution critically depends on the relative strength of policies which can permit providers: the stronger policies the provider pursues, the larger profits he gets if the policies of other providers are fixed. To see this, compare Figure 1.I where provider A_1 employs

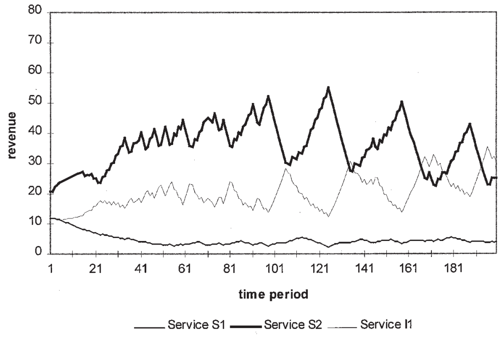


Figure 6.J

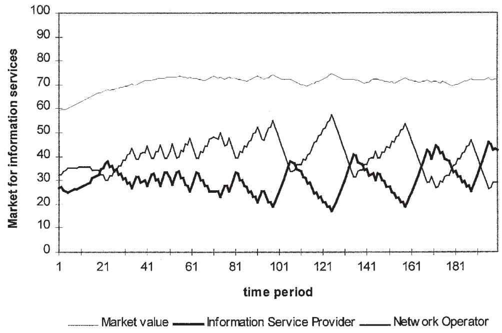


Figure 7.J

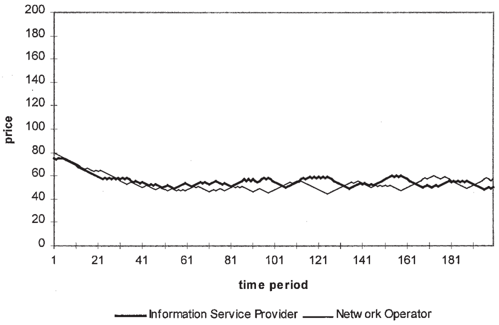


Figure 3.J

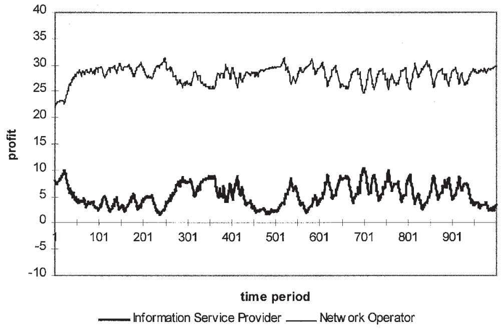


Figure 1.J.A

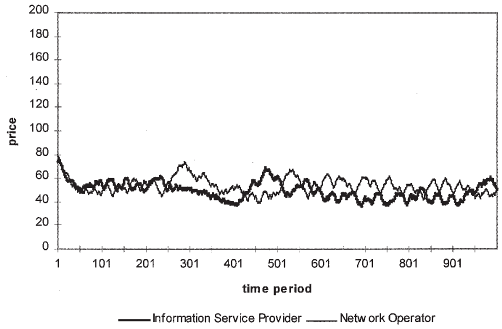


Figure 3.J.A

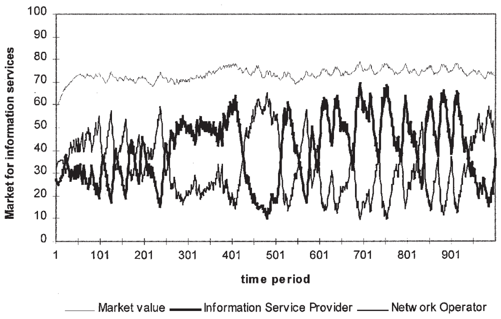


Figure 7.J.A

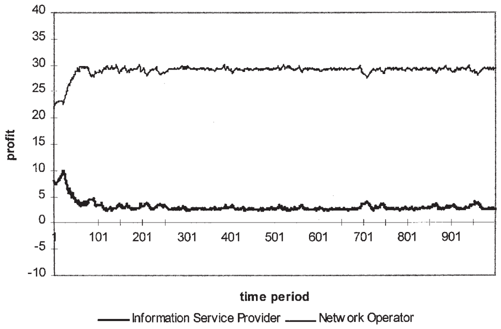


Figure 1.P

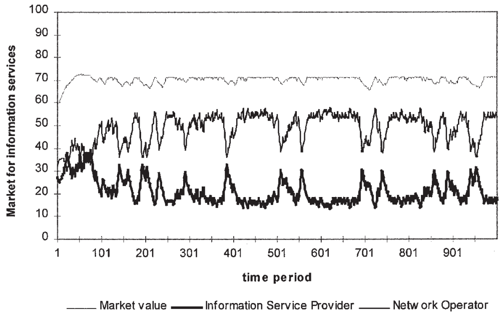


Figure 7.P

weak policies and provider A_2 employs strong policies, with Figure 1.J where both providers pursue the strong policies, and with Figure 1.K where A_1 uses the strong policies while A_2 uses the weak policies.

The case when both providers pursue strong policies merits further discussion. If we look at the revenue breakdown for provider A_1 (Figure 6.J) and at behaviour of market shares (Figure 7.J) we see that the system is very unstable and exhibits chaotic behaviour. Price dynamics show that providers are engaged in a price war (Figure 3.J). Chaotic behaviour can further be seen in Figures 1.J.A, 3.J.A and 7.J.A which differ from Figures 1.J, 3.J and 7.J by a considerably greater time span. In this case regulation (or mutual agreement) can have a strong stabilising function. The introduction of lower bound on service S_2 considerably reduces the profit oscillations (compare Figure 1.O with Figure 1.J.A). Such oscillations are reduced even more if in addition the upper bound on the price of infrastructural service I_1 is introduced (see Figure 1.P). Chaotic oscillations in market shares and prices are also reduced considerably (compare Figure 7.J.A and 7.P, Figures 3.J.A and 3.P). Figure 3.P shows another interesting consequence of price restraint: provider A_2 concentrates on the provision of higher quality services at a somewhat higher price, while provider A_1 concentrates on provision of basic services at lower price.

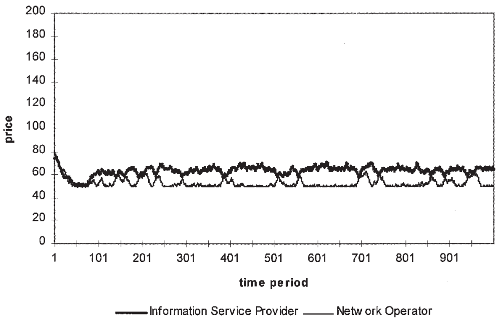


Figure 3.P

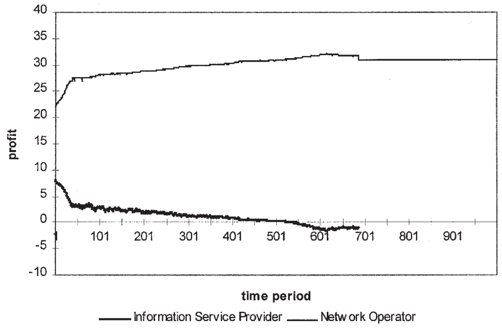


Figure 1.L

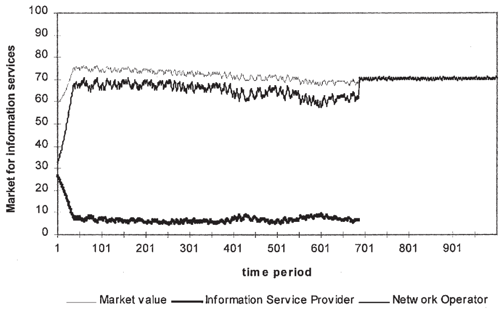


Figure 7.L

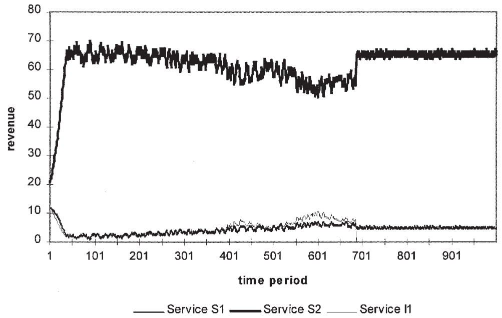


Figure 6.L

Similar to the previous case, provider A_1 succeeds in driving A_2 out of business if the policies of the latter are too weak (Figure 1.L). This, however, does not lead neither to the profit decrease for A_1 nor to the decrease of the total market value for information services (Figure 7.L). This is because in this case provider A_1 can provide the competitive service S_2 and takes over the customers of provider A_2 (Figure 6.L). Again, regulation which fixes the upper limit for the price of I_1 permits provider A_2 to survive (Figure 1.M) but this does not have any beneficial effect on the market contrary to the previous case.

Predictably, higher prices for maintaining provision capabilities lead to lower profits, but do not change the qualitative picture (Figure 1.Q without regulation and Figure 1.S with regulation).

7 Summary

We have here presented a general Agent Net methodology for modelling complex distributed multi-agent systems found in telecommunications, and presented a prototype of the simulator MODAGENT for simulation of such systems. Our reference point was its application to modelling of information industry, although it is applicable also to other multi-agent systems.

This project is in the stage of further development. Some of the issues which we address now are the following:

- As we have seen multi-agent systems exhibit widely different dynamics under different values of system parameters. The tools are necessary which would permit us to identify the regions of stability of certain equilibrium points and regions where one strategy is superior with respect to another strategy. One possibility to develop such tools is to extend to the multi-agent systems the theory of sensitivity analysis developed for Discrete Event Dynamic Systems [1, 9, 17, 21, 28, 30].
- Development of robustness notions for the design of enterprise strategies and their evaluation in a multi-player environment.
- Game theoretical approaches for strategy evaluation.

References

1 Archetti, F, Gaivoronski, A, Sciomachen, A. Sensitivity analysis and optimisation of stochastic Petri nets. *Discrete Event Dynamic Systems : Theory and Applications,* 3, 5–37, 1993.

2 Archetti, F, Gaivoronski, A A, Stella, F. Stochastic optimisation on Bayesian Nets. *European Journal of Operations Research,* 101, 360–373, 1997.

3 Arthur, W B. Bounded rationality and inductive behavior (the El Farol problem). *American Economic Review,* 84, 406–411, 1994.

4 Arthur, W B. *Increasing returns and path dependence in the economy.* Ann Arbor, Michigan University Press, 1994.

5 Birge, J R, Wets, R J-B. Computing bounds for stochastic programming problems by means of a generalized moment problem. *Mathematics of Operations Research,* 12, 149–162, 1987.

6 Bonatti, M, Ermoliev, Y M, Gaivoronski, A A. *Modeling of multiagent systems in the presence of uncertainty : the case of information economy.* Laxenburg, IIASA, 1996. (Technical Report WP-96-94.) (To appear in Journal of Robotics and Autonomous Systems, 1998.)

7 Bonatti, M, Gaivoronski, A A. Economic modelling for strategical planning and top level engineering of information infrastructures : overall approach. In: *Proceedings of NETWORKS'96, Sydney,* 1996.

8 Bonatti, M et al. *Information industry enterprize model, reference models and items to be standardized and how.* European Telecommunications Standards Institute, March 1996. (Report ETSI/EPIISG(96)16.)

9 Cassandras, C G. *Discrete event systems : modeling and performance analysis.* Irwin Publishers, 1993.

10 Dempster, M (ed.). *Stochastic programming.* London, Academic Press, 1980.

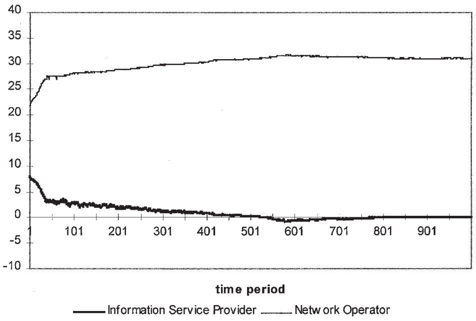


Figure 1.M

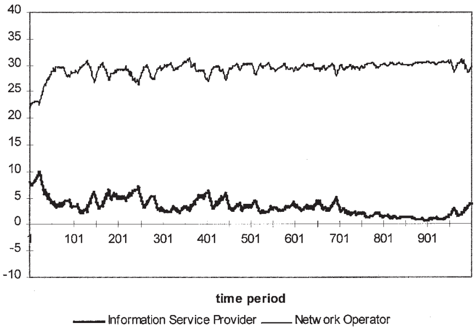


Figure 1.O

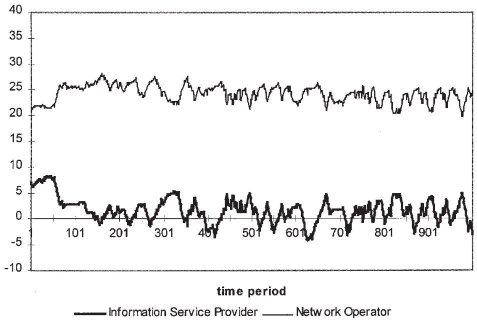


Figure 1.Q

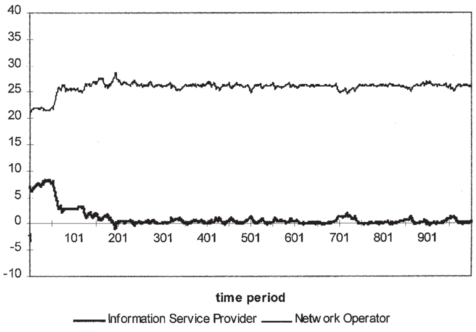


Figure 1.S

11 Dosi, G, Nelson, R. *Evolutionary theories in economics : assessment and prospects.* Laxenburg, IIASA, 1993. (Technical Report WP-93-064.)

12 Ermoliev, Y, Gaivoronski, A A. Stochastic quasigradient methods for optimization of discrete event systems. *Annals of Operations Research,* 39, 1–39, 1992.

13 Ermoliev, Y, Wets, R J-B (eds.). *Numerical techniques for stochastic optimization.* Berlin, Springer, 1988.

14 Even, R, Mishra, B. *CAFE : a complex adaptive financial environment.* New York, Courant Institute of Mathematical Sciences, 1996. (Technical report.)

15 Gaivoronski, A A. Approximation methods of solution of stochastic programming problems. *Cybernetics,* 18, 97–103, 1982.

16 Gaivoronski, A A. Convergence properties of backpropagation for neural nets via theory of stochastic gradient methods. Part I. *Optimization Methods and Software,* 1994.

17 Gaivoronski, A A, Shi, L Y, Sreenivas, R S. Augmented infinitesimal perturbation analysis : an alternate explanation. *Discrete Event Dynamic Systems : Theory and Applications,* 2, 121–138, 1992.

18 Gaivoronski, A A. *MODAGENT, users manual.* Trondheim, Norwegian University of Science and Technology, 1998. (Technical report.)

19 Hertz, J, Krogh, A, Palmer, R G. *Introduction to the theory of neural computation.* Redwood City, Addison-Wesley, 1991.

20 Higle, J L, Sen, S. Stochastic decomposition : an algorithm for stage linear programs with recourse. *Mathematics of Operations Research,* 16, 650–669, 1991.

21 Ho, Y C, Cao, X R. *Perturbation analysis of discrete event dynamic systems.* Boston, Kluwer, 1991.

22 Kall, P, Wallace, S. *Stochastic programming.* New York, Wiley, 1994.

23 Lane, D A. Artificial worlds and economies. *Journal of Evolutionary Economics,* 3(2 and 3), 89–108 and 177–197, 1993.

24 Mulvey, J M, Ruszczynski, A. *A new scenario decomposition method for large-scale stochastic optimization.* Princeton Unversity, 1992. (Technical Report SOR-91-19.)

25 Neapolitan, R E. *Probabilistic reasoning in expert systems.* Wiley, 1990.

26 Nelson, R, Winter, S. *An evolutionary theory of economic change.* Cambridge, MA, Belknap Press of Harward University Press, 1982.

27 Peterson, J L. *Petri net theory and the modeling of systems.* Prentice-Hall, 1981.

28 Pflug, G. Gradient estimates for the performance of Markov chains and discrete event processes. *Annals of Operations Research,* 39, 173–194, 1992.

29 Plambeck, E L et al. Optimizing performance functions in stochastic systems. *Mathematical Programming, Series B, Approximation and Computation in Stochastic Programming: State of the Art, 1993,* 1993.

30 Rubinstein, R Y, Shapiro, A. *Discrete event systems : sensitivity analysis and stochastic optimization via the score function method.* Wiley, 1993.

31 Steiglitz, K, Honig, M M, Cohen, L M. A computational market model based on individual action. In: Clearwater, S (ed.). *Market-based control : a paradigm for distributed resource allocation.* Hong Kong, World Scientific, 1995.

32 Waldspurger, C A et al. Spawn : a distributed computational economy. *IEEE Transactions on Software Engineering,* 18, (2), 103–117, 1992.

33 Wellman, M P. A market oriented programming environment and its application to distributed multi-commodity flow problem. *Journal of Artificial Intelligence Research,* 1, 1–23, 1993.

Alexei A. Gaivoronski (45) is Professor of Operations Research at NTNU, Trondheim. He has an M.Sc. from Moscow Institute for Physics and Technology and a Ph.D. from V. Glushkov Institute of Cybernetics, Kiev, both in applied mathematics specialising in operations research and optimisation. He has worked in both academia and industry, and his scientific interests are optimisation, modelling of stochastic systems, manufacturing and telecommunication models. He has published more than 50 papers on these topics in international journals and books.
e-mail: Alexei.Gaivoronski@iot.ntnu.no

Technological Evolution: Some Strategic Considerations

JAN A. AUDESTAD

1 Introduction

The strategic problems we are facing today in telecommunications are all owing to the enormous growth of complexity [1]. The complexity is present in all aspects of telecommunications. Below we will give an analysis of the technological evolution and its consequences on markets and research: how it is now and how we expect it will change during the next few years. The analysis will comprise the existing products, the new content market, sensor technology and, finally, nanotechnology.

Some strategic problems in this context are

- How much resources should we allocate to further evolution of traditional products?
- Why do we not earn money on the new content rich services?
- Will sensor technology, nomadic computing and wearable computing create new opportunities for telecommunications?
- If nanotechnology ever materialises, will it have any impact on telecommunications?

2 Traditional Products

[2] contains one way of assessing the portfolio of products. The method consists in comparing each product in terms of four variables:

- Current margin created by the product;
- Expected short term growth in revenue;
- Current total revenue created by the product in the market;
- The market share the company has of this revenue.

Products may then be viewed as belonging to four categories [9]:

1 'Cash cows' or products with high margins but with small or no growth potential;

2 Products with low margins and little or no growth potential;

3 Products with low margins and much expected growth;

4 Products with high margins and much expected growth.

Plain telephony is an example of a cash cow. In developed countries the market is saturated and very little growth is expected. This market is, however, difficult to assess for several reasons:

- The telephone network carries services which are not just plain telephony; examples are telefax and Internet services (at least in local areas), and transaction services which also use the telephone network as carrier;
- Some of these services are fading out such as telefax, since telefax is replaced by e-mail;
- Some of the services using the telephone network are increasing such as Internet services;
- Services which were supplementary, are now substituting each other – mobile telephony is a good example here;
- Growth in mobile telephony causes secondary growth in telephony because one of the persons connected is generally located in the fixed network.

From this it is not easy to interpret the evolution of the market for telephony by only measuring the traffic on the network; we must also differentiate between the different usages of the network. The reason for this is that, by only measuring the traffic, it is not evident how money should be used for investments and research: investments in plain telephony technology or in IP networks, or research in basic switching technology or in new types of sources such as sensors?

One likely option here is to invest as little as possible both in money and intellect on plain telephony, and rather direct the efforts towards the new opportunities. As I will explain in Chapter 3, it does not really matter on which network new services are transferred as long as the service meets the expectations of the market.

Plain telephony is a cash cow now but competition, and in particular regulation, may lower the prices and reduce the margins unless the cost is reduced. This may require a novel and innovative way of allocating costs between different usages of the network.

Plain telephony and infrastructure products like access and transport may – because of price pressure – move from category 1) to category 2); that is, become products with low margins and little or no growth. From a business point of view, such products should be terminated. These products may then become a real problem for the operator: the products must be retained for societal reasons, and the products will still represent a major part of the overall telecommunications market, perhaps 70 % or more. The size market does not matter as long as there is nothing to earn in it.

If the company has other products in category 2) without such constraints, they should be terminated before they cause losses to the company unless they create a secondary stream of income. This is an important consideration for all the products.

Products in category 3) are particularly dangerous because they may indicate the possibility of wrongly priced products or that the cost is too high for these products. There are many indications that several of the Internet products are in this category where we by Internet products mean all the individual products for processing, storing, marketing, billing, maintaining and operating content rich services. Two related problems are that the income may come through different market channels (for example, creating more demand for ISDN installations), and that this cross-subsidisation may be against regulation. In more extreme cases the revenues may come via apparently unrelated channels such as sales of energy. Telecommunications may then be delivered at very low prices because the infrastructure is also used for remote control of the energy delivery.

Mobile communication has been a product of category 4). Because of oligopolic competition with focus on price wars and new roaming conditions for operators without their own networks, this product may move towards category 3); that is, maintaining a high growth while the margin is going down.

The strategic problems are then:

- To decompose products into parts which may be products on their own – such as in the example of telephony explained above – and to analyse each such part separately;
- To identify to which categories the different products of the portfolio belong;

- To evaluate how competition and regulation may move products from one category to another, and to assess the impact of such changes;
- To assess the impact of products of category 2) if they are required to be retained by society.

3 Content Rich Services

Many content rich services are part of the existing portfolio of telecommunications operators. However, as I see it, the understanding of such services is much related to the problem of dividing the main product into parts. This may be done as shown in Figure 1. In the existing context the content rich services are related to the Internet. This may change; however, the general concept of Figure 1 is likely to persist.

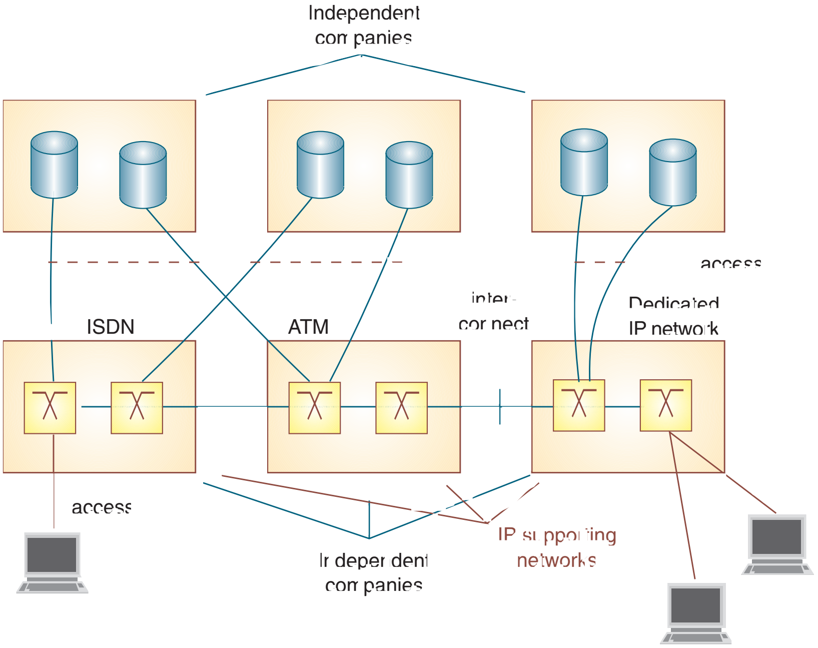


Figure 2 Technical arrangement

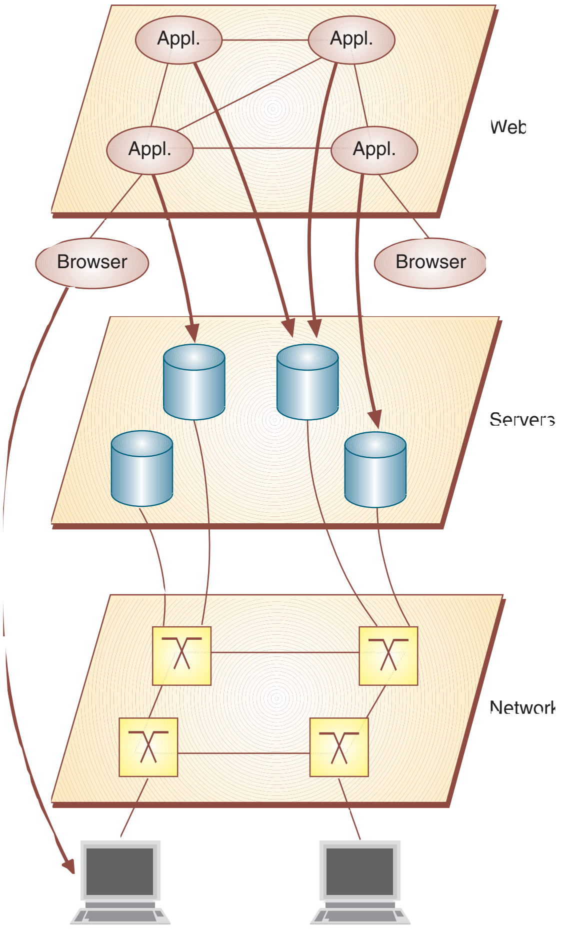


Figure 1 Model for Internet

In order to understand the services we divide them into three separate entities: the web, the Internet servers, and the supporting network. These three layers are independent in the sense that the technology may be replaced independently and at different rates in the three layers. The different layers also represent separate services or sets of services.

The Internet uses several types of networks for transporting bits: telephone networks, ATM networks, mobile networks, cable TV networks, and so on. There may be several networks of the same or different type in tandem. The only requirement is that these networks interwork with one another without destroying or changing the bits transferred from one end to the other. The terminals are connected via some type of access to the network. There is no requirement concerning the type of terminal to be used except that it must support the browser software which is the 'terminal' as seen at the application level.

The networks also interconnect the servers which are computers designed for running multiple applications in a distributed environment. The servers are connected to the networks via access systems. The servers communicate by using the Internet Protocol (IP). For simplicity I have not shown additional equipment such as routers required for supporting this communication. The operation and management of servers are businesses independent of network operation. Like networks, there may be a large number of companies in the business of offering servers (and routers and other facilities). What is important is that ownership, location and operation of servers are independent of the supporting network infrastructure. The technical arrangement is shown in Figure 2. This architecture will be stable for a long time: the cost of replacing the network infrastructure is too high. This also applies if dedicated IP networks are introduced.

The application (or web pages or software programs) are shown as a separate plane. This is so because the applications are separate businesses: electronic newspapers, electronic trade, information services, education, office support, and so on. These companies own the application but they may outsource the processing and the management of the application to companies owning servers, or they may own their own servers. In the latter case they are present in two businesses at the same time: applications and servers. The

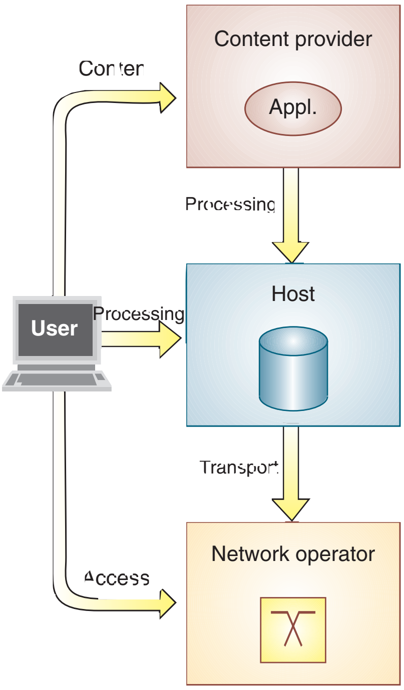


Figure 3 Customer relationships

model is not changed by this or if some company owns all three layers.

At the application level addressing is done by hypertext pointers. The server maps the hypertext pointer to an IP address which is again mapped to telephone numbers, mobile subscriber numbers or whatever numbering system is used by the network.

Generally, we may then have a structure as shown in Figure 3. The user is a customer of the network operator (access). In addition the user may be a customer of the host (owner of servers) and the content provider. The content provider is a customer of the host, and the host is a customer of the network operator.

The payment for using the applications may be made to the network provider who pays the host and directly or indirectly the content provider; or the customer pays the content provider who pays the other operators. The customer may also pay all three independently. The payment method may depend on the application.

Note also that content providers may co-operate in delivering a joint service.

One problem with the Internet is that it is difficult to earn money on it. There may be several reasons for this:

- The co-operation between hosts and network operators may be such that the host does not receive enough money from the network operator in order to be adequately paid for the value created;
- It may be the other way round: the price paid by the host for using the network may be too small;
- The use of the network may be too expensive for the host and the content provider so that they lose money on the service;
- There may be much arbitrage where the content provider and the host use networks paid for by others, such as governments; the 'old' Internet consisted of mainly 'free' circuits;
- The market mechanisms are really not understood: the Internet may be a service offer to people without money but creating much communications which means that prices accepted by the market are low.

Another problem with content is that the service may simply not be good enough. Take electronic trade as an example. If you know exactly what you want, it is simple. If you do not, it is not so easy. On one screen we can only show a small fraction of the goods you may see by strolling through a shop. Take the "world's largest bookstore" – the Amazon database. They claim to have two million titles, which is three times more than the biggest physical bookstore. If you are just going to browse through such a big store, it will take time. Suppose that you can read one book title and synopsis every minute, then the browsing takes 30,000 hours. If you spend 12 hours every day on this endeavour, it will take you six years to finish the browsing. Then you may just as well start again because many of the titles are either obsolete or new!

This problem of electronic trade is purely psychological and not technical. It is easy to make the technology: virtual reality shopping malls, huge databases like the Amazon, and customer specific advertisements and construction of the mall (where shops may be shown in accordance with your habits and interests learned by earlier visits to the mall).

It is also difficult to enter many of the content businesses. One reason is that they may require co-operation completely different from other systems. This implies also ownership of the business. In [1] the C4 convergence was explained. In the early phases of this development it was assumed that new companies or at least firm alliances would form along the C4 axis content, computers, communications and consumer electronics. So has not happened [3]. The reasons seem to be that such vertical integration is not economically optimal for each of the actors along the axis, and that the need for new co-operation changes faster than the time required to build and stabilise a new company or alliance. In the most complex cases companies may co-operate in one area and compete in others.

This type of dynamic market, competition and co-operation has not existed before and is new to the participants in the market. Traditional administration and value monitoring does not apply to such systems. I believe that this is the main reason why the content market has not materialised.

4 Sensors, Wearables and Nomadic Computing

The Internet as a commercial product is only five years old. Sensor technology, as we shall describe it here, is even younger. Of products now coming are video cameras, radars and navigation receivers on one chip. The video camera is complete with lens and radio communication. The navigation sensor can be sewn into clothes and utilise the GPS standard for determining position. The radar can be used for automatic measurement of the oil level in motors and the content of tanks.

The artificial nose is under development. Still it is not at microchip size but that will come soon. Already gass chromatographs are down to this size with the channel etched into the semiconductor which also contains the sensing device, the spectral analyser and the output formatting device. Microprocessors may replace the bar code on merchandise. The electronic handcuff monitors people sentenced to mild imprisonment with some restrictions on their mobility. The list of possible devices is increasing day by day [4].

Some of these devices are wearables and wearable computers; that is, devices attached to the body. Some of them may even be placed in the body or eaten (edibles) for medical analysis and monitoring.

The combination of wearables and computer networks is called wearable computing [4]. This is again part of what is called nomadic computing. Some characteristics of such systems are:

- The core device is a network of computers;
- They may contain sensors of some kind (GPS receiver, video camera, thermometer);
- Some or all of the computers may be attached to a person, car, container or other mobile 'platform';
- The computers may offer interfaces which are context dependent (keyboard, audio, screen) and which may change as the context changes;
- They may require much communication in web-like structures and on radio interfaces;
- Their applicability may be very sensitive to the cost of communication.

Some of these technologies are not new; they are only reappearing in a new form taking advantage of the miniaturisation. Early systems were tracking devices on animals and birds, navigation stations monitoring movements of glaciers, remote weather stations and pollution monitoring and flood warning systems. These systems had several constraints such as price, size and lack of computational power. The new sensor technology is making some of these technologies ubiquitous, and enabling new applications to emerge.

The strategic problems for the telecommunication industry are:

- Which of these applications will really take off?
- Do the sensor technologies require new network technologies?
- How much traffic will they generate?
- How will the traffic be distributed over the network?
- How should such traffic be priced in order for the applications to materialise?

5 Nanotechnology

The idea of nanotechnology is to make low cost manufacturing technologies on the atomic level. Part of this technology is self-reproducing systems and assembler devices [6], [7]. Some well-known results are producing the IBM logo at atomic level made from 35 atoms of xenon; AT&T's 'three little gears'; and the molecular sleeve bearing. The idea of such systems came up around 1970. In 1992 the US Senate got interested in the idea because of the enormous impact the technology will have on society – if it is realisable [8].

If efficient assembler systems – like the ones that exist in cells – can be made, the technology offers opportunities like making food from garbage, producing cancer cell killing robots of the size of a blood cell and processors with molecules as components. It is claimed that these technologies will be available sometime during the next ten to thirty years.

As we saw above, systems offering access to information such as the Internet create much telecommunication traffic. The sensor technology requires applications on fixed and nomadic platforms to be connected together possibly creating even more traffic. The communication requirement coming as a result of the nanotechnology may be even much bigger.

The point of bringing in such far-fetched technologies is that the commercialisation of the World Wide Web took us by surprise. We are just getting in touch with the development of the sensors.

	Traditional technologies	**Internet technologies**	**Sensor technologies**	**Nano-technology**
Maturity	Full	Infancy	About to be born?	Embryonic?
Business potential	Stagnated/ declining?	Big?	Enormous?	Astronomic??
Knowledge and mastery	Well-known, fully mastered	Partly understood, not mastered	Not understood	Not understood
Current research focus	Almost all	Little	Almost none	None
Required degree of research focus	Little	Much on applications, little on technology	Much on applications, much on interconnectivity, much on technology	Some: at least know what it is and how it develops. Must have some degree of preparedness

Figure 4 Five strategic imperatives and existing and future products

Again the telecommunications industry may be unprepared. Nanotechnology is a natural evolution of these technologies and must therefore be taken seriously so that we at least know what is going on.

6 Strategic Imperatives

This leads us to Figure 4. Here we have shown five important aspects of the four families of evolution we have discussed above.

The most important point is that most of our focus today is on the traditional technologies. In this area we – that is, the whole telecommunications industry – direct almost all research efforts. Very little is directed towards research on Internet applications. Still much of the research money in this area is on network and protocol issues. Very little effort is put into sensor technology and the impact this technology may have on our systems.

Now the time has come to shift this focus, where most of the new research efforts should be toward Internet applications and sensor technology. The traditional technologies are more and more becoming cash cows, and as such we should avoid wasting our efforts on bringing the technology too far forward. In the current situation we cannot anymore afford to spend fifteen years on developing a technology like ISDN which is obsolete when it is introduced (in a funny way, the Internet demand for more bandwidth has resulted in an unexpected demand for ISDN). The reason why we should shift the attention in this way is that the new technologies will have enormous impact on what we are going to earn money from and how to conduct our business in the future.

References

1 Audestad, J A. Telecommunication and complexity. *Telektronikk,* 94, (3/4), 1998, 2–20. (This issue.)

2 Audestad, J A, Kjæreng, A, Mahieu, L. An object oriented model of the telecommunications business. *Telektronikk,* 94, (3/4), 1998, 54–61. (This issue.)

3 Baldwin, T, McVoy, D S, Steinfield, C. *Convergence.* Sage Publications, 1996.

4 Mann, S. *'Smart' clothing.* MIT Media Lab (http://www.wearcom.org/smart_clothing/)

5 Mann, S. *Definition of 'Wearable Computer'.* Presented at 1998 International Conference on Wearable Computing ICWC-98, Fairfax VA, May 1998.

6 Merkle, R C. Self replicating systems and molecular manufacturing. *Journal of the British Interplanetary Society,* 45, 1992, 407–413.

7 Merkle, R C. Design considerations for an assembler. *Nanotechnology,* 7, 1996, 210–215.

8 Regis, E. *Nano.* Bantam Books, 1995.

9 Audestad, J A. How to survive in the future. *Telektronikk,* 94, (3/4), 1998, 74–81. (This issue.)

Jan A. Audestad (56) is Senior Advisor for the Corporate Management of Telenor. He is also Adjunct Professor of telematics at the Norwegian University of Science and Technology.

e-mail: Jan-Arild.Audestad@s.hk.telenor.no

An Object Oriented Model of the Telecommunications Business

JAN A AUDESTAD, ARVID KJÆRENG AND LAURENT MAHIEU

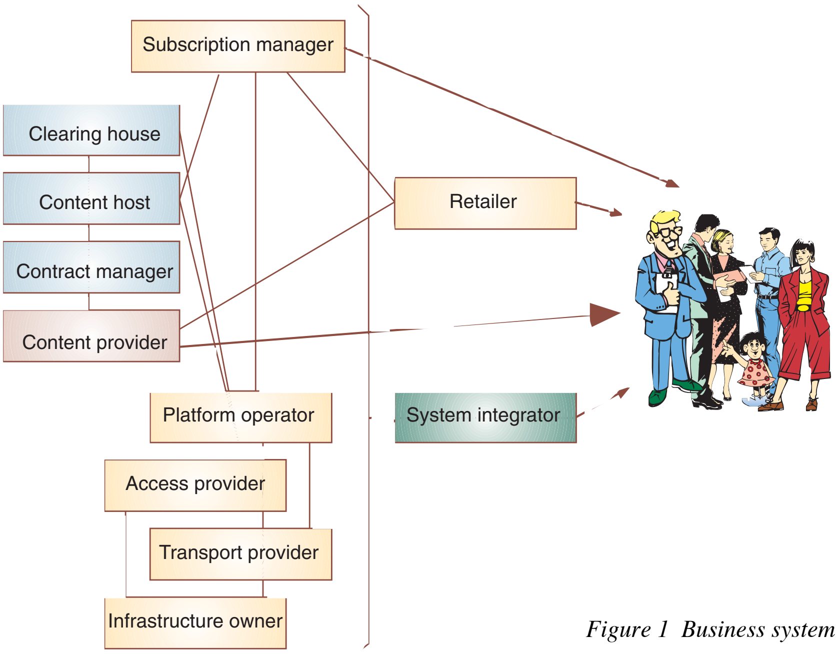


Figure 1 Business system

1 Introduction

We saw in [1] that the organisation of telecommunications is becoming very complex. The business may best be described by decomposing it into several roles. For completeness and easy reference the model is reproduced in Figure 1. Each box in the model is called a *role* which may correspond to a business area of a bigger company or be a company proper. In the liberalised telecommunications market more and more combinations are seen. This is also a result of the complex delivery of services beyond simple person-to-person communication (Internet, transactions, outsourcing).

We saw also in [1] that the telecommunications business consists mainly of value networks with both vertical and horizontal co-operation (or competition) as shown in Figure 2. In the figure several companies are shown; some consist of many roles and some only of one role.

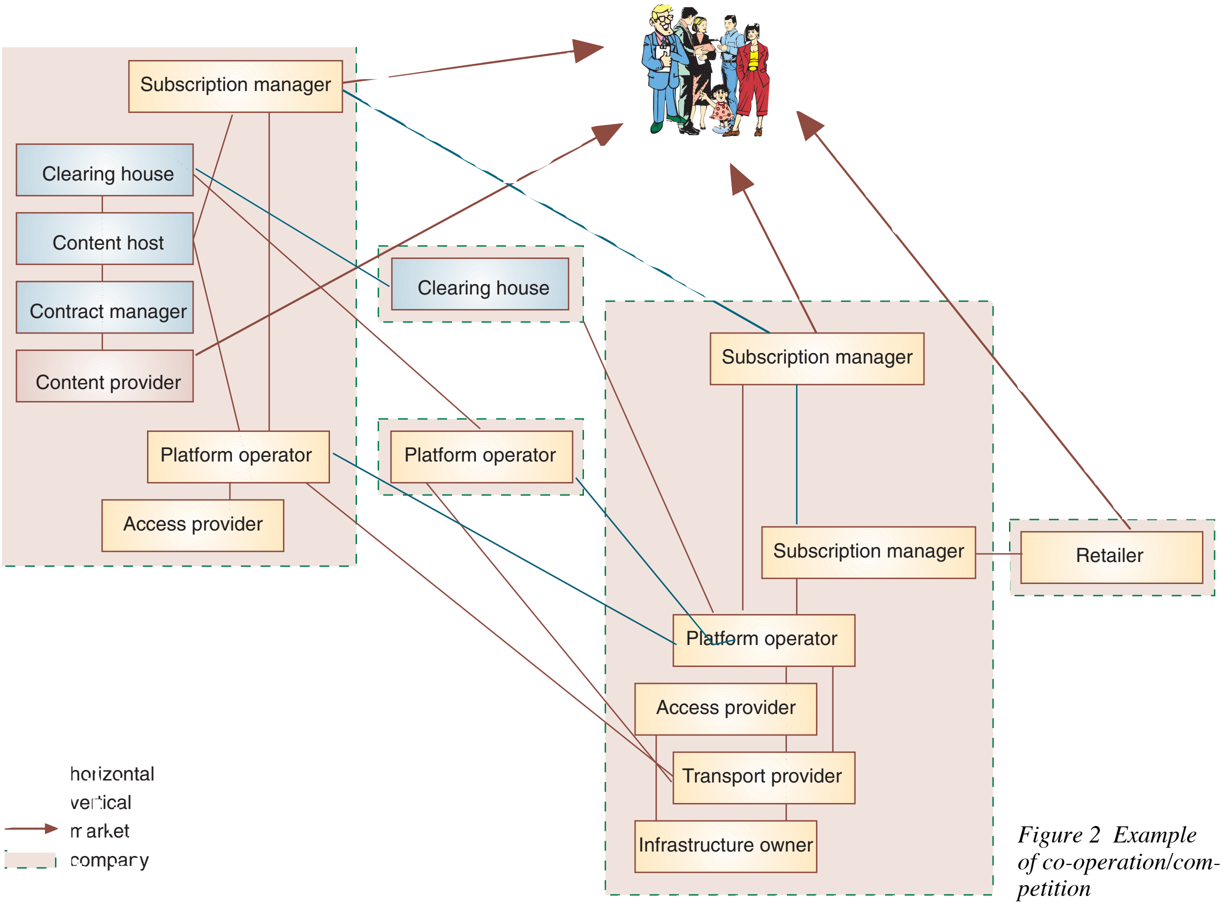


Figure 2 Example of co-operation/competition

These companies co-operate or compete in delivering products to the market. We say that there exists a *relationship* between the roles if they co-operate or compete. There are two types of relationships: *vertical* between different roles, or *horizontal* between similar roles. The notion of vertical and horizontal relationships is independent of whether the roles are in the same or in different companies. The difference between internal and external relationships are only qualitative, and the regulation ensures that if one role is within the incumbent, the relationship is the same independently of whether the co-operating role is within the same company or belongs to a competitor.

The business structure describing all operators within a given area is complex. Figure 2 is only a simple example compared to what reality may look like. In order to understand such complex structures methods must be developed which hide complexity not relevant to the problem but contain all information required for understanding the types of interaction and the content of it. The method proposed here is built on object orientation.

2 Principles of Object Orientation

Object orientation is mainly a tool for software construction and is usually associated with a programming language: Simula, C++, Eiffel and Smalltalk. In definition of Open Distributed Processing (ODP) defined by the International Organization for Standardization (ISO), object orientation is used also for definition of systems in such a way that the specification is not directly transformable to software. This is the way in which we will use the method here.

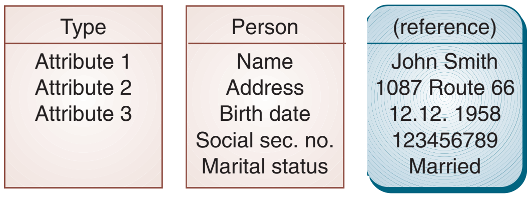
An *object* represents a concrete thing. If object orientation is used to describe road traffic, then cars, pedestrians, roads, pedestrian crossings, junctions and traffic lights are all *objects* defined by their behaviour. The behaviour of the object is confined to the case where it is used, and consists of the definition of what it is composed of (*attributes*) and *operations* (or *methods*) which can be used in order to read or change one or more of these attributes. The description contains only such attributes which are relevant for the problem at hand. In the example of road traffic, where the focus is to determine how queues are created, a car may be fully defined by its speed, direction of travel and location. For this problem it is irrelevant how big the motor is, how many passengers the car carries, the size of the petrol tank, and what type of petrol injection system is used. Similarly, the traffic lights are defined by which colour is shown and the cadence for shift between different colours. In this way information not relevant is suppressed in order to make the system easier to understand. Moreover, the only thing that the surroundings know about an object are the attributes and their current value (or, what is equivalent, the *state* of the object). It is not shown how the object changes state, that is, all such algorithms as well as other internal constructions are hidden.

If we take the windows system of a computer, the attributes of a window object may be all controls available for operations on the window (file management, editorial functions, size and location management, acknowledgements, terminations, help functions and so on). The operations are everything the user may do with each attribute (use the mouse to move a window or request a given action, or to enter a given set of characters). How these operations are performed and which algorithms are used are not visible to the user. This is one of the most important advantages of object orientation: *information hiding*. This mechanism enables us to distinguish between what is publicly known about an object and what is private. The *public view* consists then of the attributes and how we can interact with them. This is what the user of the object is allowed to see, and contains everything required for using the object. In Windows this corresponds to information contained in the user manual. The *private view* is all software, hardware and other elements required to make the object perform its tasks. In Windows this corresponds to the software construction of the system and is found in specification manuals in the company implementing this particular product.

The second important aspect of object orientation is *generalisation*. Every person is an object. For a given application every person object contains the same attributes, for example, name, address, date of birth, social security number and marital status in public registers; or name, job type, skill set and performance rating in a personnel database of a company. A person described in both these ways are described in terms of different objects. However, every person contained in the public register is described in the same way; the same applies to every person contained in the personnel database. Objects described by the same set of attributes are said to belong to the same *object type*. All persons contained in the public database are of the same type. Each individual person is described by a specific name, address, date of birth, social security number and marital status. We then say that each individual person corresponds to an *instance* of the object type. The concept of generalisation makes it possible to describe a large number of individual objects once only in term of the type from which it is instantiated (or brought into life if it is an object that can be implemented in software).

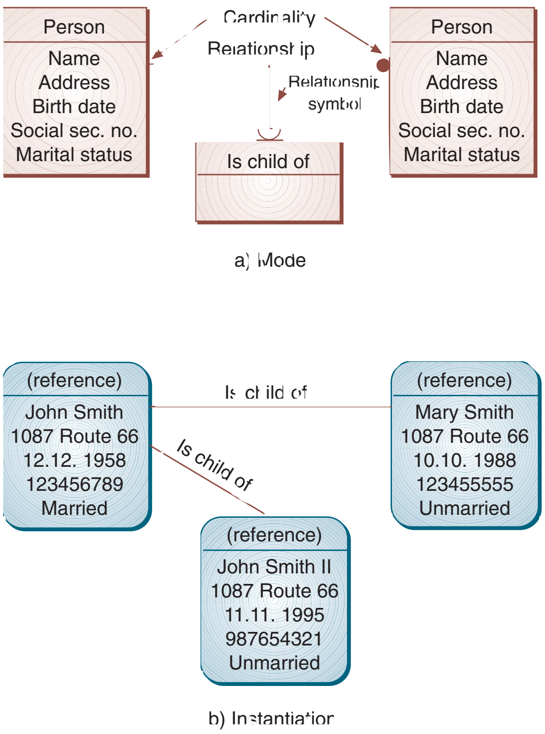
The third important aspect of object orientation is that, in models, *objects with exactly the same attributes – that is, are of the same type – are equal and replaceable*. This is one of the consequences of the generalisation principle. In practise, this may not always be possible; a windows program may not be adaptable to a given operating system of a computer because of constraints in the part of the object which is invisible to the user. This has to do with construction practises and monopolisation of products, and has nothing to do with the principles of object orientation. In the models we make we will always assume that exchangeability of objects is possible.

Graphically we may illustrate an object as a rectangle as shown in Figure 3 [2]. The object is then defined by its *type name* and its attributes. The operations we may do towards the attributes are generally of two types: we may read the value of the attribute, or we may change its value. For certain attributes it may not



a) Type definition b) Example of type c) Instance

Figure 3 Graphical notation of objects



a) Model

b) Instantiation

Figure 4 Example of relationship

be allowed to change the value of it, that is, the write operation is not allowed. As an example the object type person of the public database is shown. In this object type it may not be allowed to use the write operation against birth date and social security number. The instance is also shown. Note that the instance may be distinguished from the type by depicting it as a rectangle with rounded corners.

Objects do not stand alone in a model: they interact with other objects. This interaction is called a *relationship*. The relationship may also be complex and contain attributes, that is, the relationship is also an object. Like objects, relationships may be generalised into *relationship types*. Relationships may be depicted as shown in Figure 4. Here two objects of the same type (person) are related by the relationship *Is child of*. Note that we have introduced another term called *cardinality*. The cardinality is the number of objects a given object may be related to. In this case the cardinality of the person object on the left is one (every child must have only one father or mother); the cardinality of the object to the right (the child) is arbitrarily many, indicated by a filled circle, since a person may have none, one or any other (reasonable) number of children. The figure also shows an instance of the model, where John Smith has two children. There are other symbols for cardinalities: more than one (see Figure 5), exactly *n* (indicated by the number *n*), zero or one (open circle).

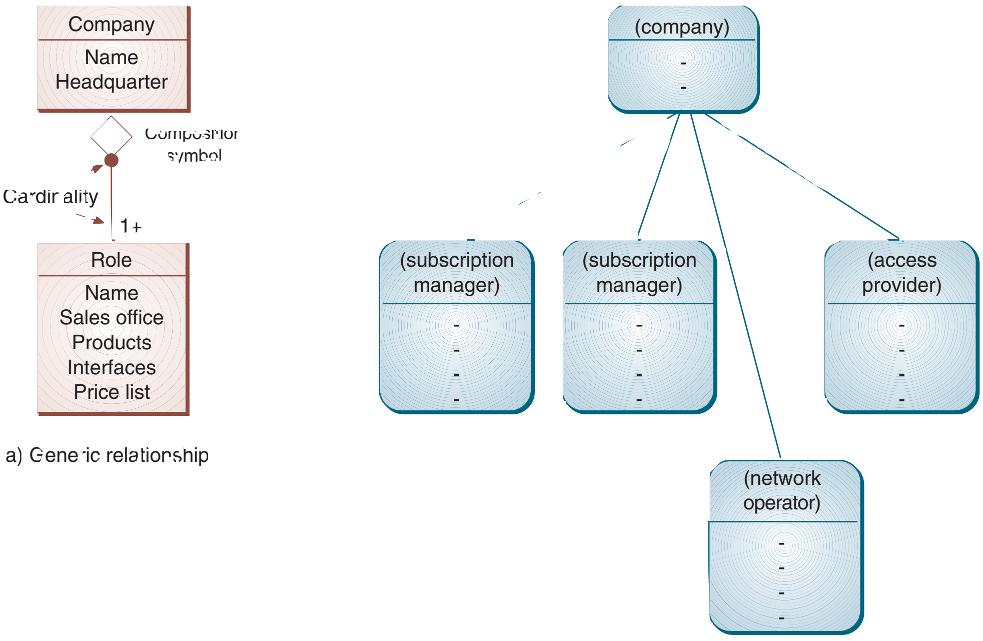
A fourth important property of objects is *composition*. This means that an object is composed of several other objects called *components*. The object itself is called a *composite*. Composition is also a relationship type. Since it is generic, it may be convenient to use a special symbol for it as shown in Figure 5. In this example we indicate that a company may consist of several roles taken from the business system. The company may then be instantiated as shown. Note that the instance may consist of several roles of the same type, for example, subscription managers delivering subscriptions to different market segments. The 1+ cardinality against the role object type indicates that the company must contain at least one role in order to be defined as a company.

Now we are prepared to look at the business system in more detail.

3 Example

In Figure 6 we have shown a small part of the business system of Figure 2 in object oriented notation. The model has apparently not become simpler. When identifying the relationships this is usually what happens. However, in order to manage the model and find all relationships, the model may be simplified. First we may delete the company objects and the composition relationship. This has nothing to do with the business relationship; this is organisation of the company. Next we may split the model into independent parts where independence means that there is no direct relationship between the parts of the model. Doing this, we have to be very careful so that relationships are not overlooked.

Figure 7 shows an example of this decomposition. Here we have decomposed the model into three submodels: sales, management and interconnect. By some training this way of handling the model becomes simple, and we obtain models which show exactly what we want and from which we may draw important conclusions.



a) Generic relationship

b) Instantiation

Figure 5 Composition

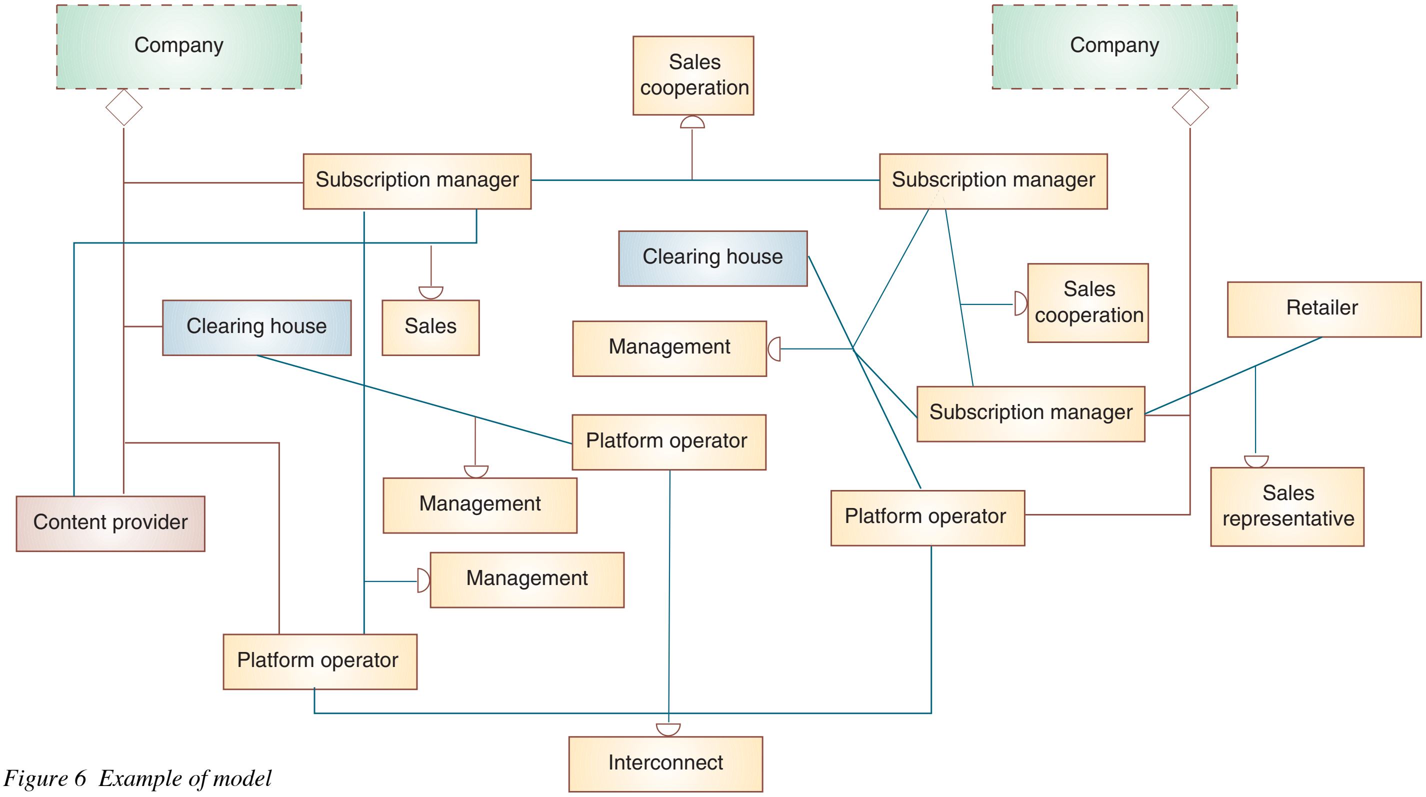


Figure 6 Example of model

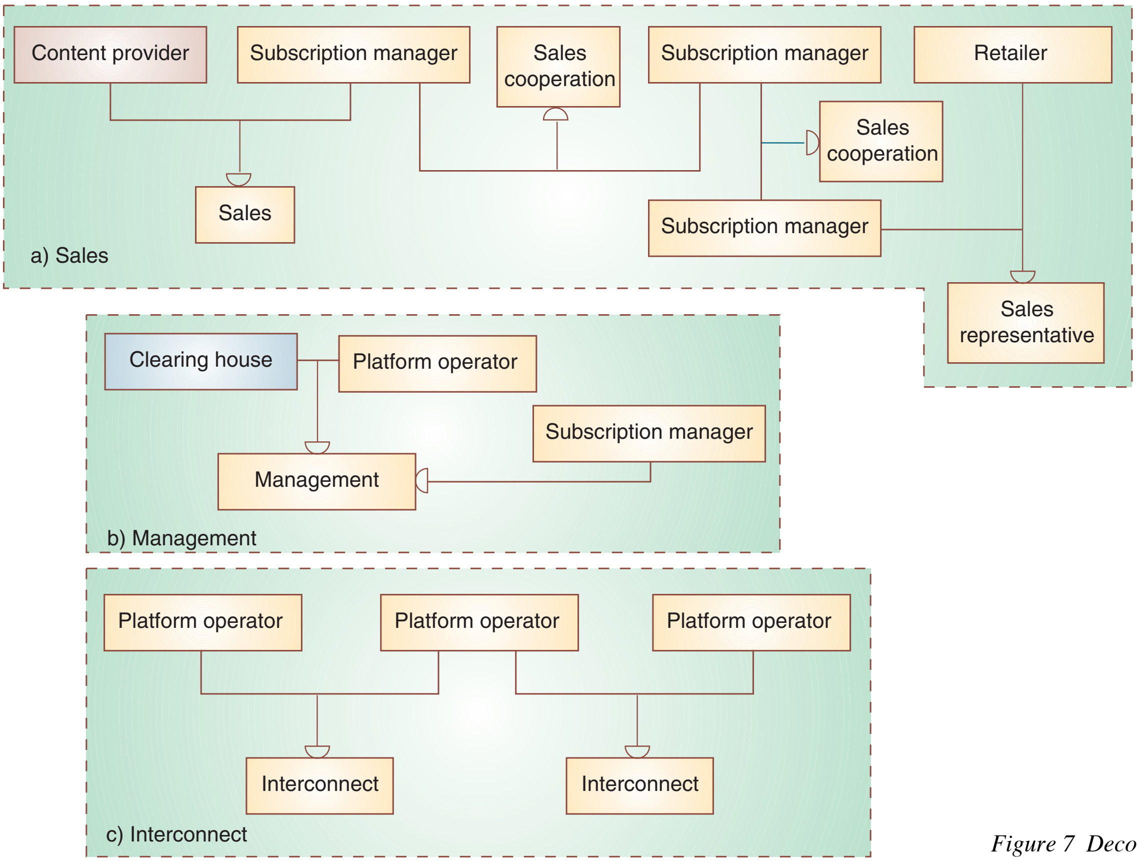


Figure 7 Decomposition

4 How to Identify Objects and Relationships

Here are some methods by which we may identify objects and relationships in a particular application of telecommunications. Note that it is far too general to make an overall model of telecommunication businesses. The basic relations are found more easily for concrete examples. However, from such an analysis we will find that the same relationship objects are appearing several times. This means that it is not only for object types the model splits into a rather small number of distinct objects. The same applies for the relationships.

Some useful methods are as follows. The object types are those contained in the business model in Figure 1.

1. Decomposition of the company is the starting point in the analysis, that is, to identify which roles it consists of and which attributes are implemented in the roles. This may uncover if there are attributes lacking or being wrongly defined, or if there are too many attributes showing information about the company which should be hidden. Note that the attributes are what we will let the environment know about the role. This may also uncover if the company is not complete with regard to role composition.
2. Other roles which will not be implemented in the company must then be identified in order to see how we can include these roles in our business by alliances. The alliances may be internal or external. This is one way of completing the company with regard to roles and attributes since attributes are also obtained by alliances.
3. The company will make products where some products can be made entirely within the company while other more complex products will require that many roles co-operate in the production. Some of these products may require co-operation far beyond the role definition defined above, and may in themselves require extended models where for instance the content provider may be decomposed. For each product it is, therefore, necessary to define which roles are involved and how these roles are divided among internal and external relationships. Detailed analysis of how the roles interact should be part of this exercise.

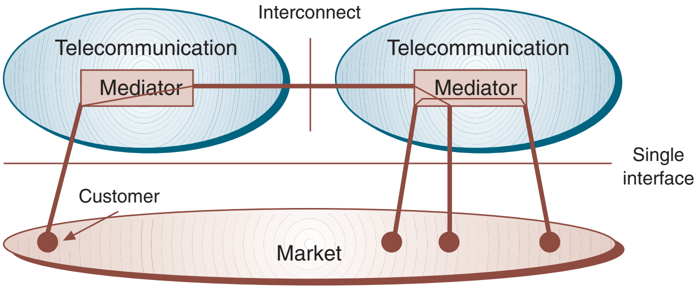


Figure 8 Symmetric mediation

4. The attributes of a role object define which interfaces other roles will have with that object. The activation of attributes will also depend on what type of alliance or competition takes place with other roles. This means that the interface may be different depending upon where and what a co-operating role is.
5. This leads to a more general analysis of alliances: which roles each partner should have, roles shared between partners and roles existing independently in each partner.

These are analyses of how a given telecommunication company performs its business alone or in alliances. Concerning competitors similar considerations apply:

6. It is particularly important to identify the role composition of the competitor and possibly how it may evolve. This also involves analysis of which products such a role composition may produce.
7. The competitors may also form alliances.

For the incumbent and for newcomers in the market it is important to understand the regulation of the market. The regulation will be different for the two types of actors.

8. In the business process not all interfaces may be regulated nor which conditions apply at the interface. On the other hand, interfaces which are internal in one company may be demanded opened by the regulator for forced co-operation. In this respect the problem is to identify what are equal conditions. The analysis should include such things as possibility and restriction of arbitrage, risk imbalance and business denial. Part of the analysis should be concerned with how products are complementary, supplementary, or replace each other. This is certainly a dynamic process and may not match a static regulation.

5 Value Creation

Most of the telecommunication business consists in mediating between different customers. Principally, there are three distinctly different ways in which this mediating activity takes place.

Symmetric mediation

The market for telecommunications was symmetric. This means that the communications were between entities of the same sort: users talking to each other over the telephone, machines interacting in order to complete transactions, and sending of documents such as e-mail and telefax from one machine to another. This gives a mediating structure as shown in Figure 8. There is only one interface towards the market and only one market. This type of mediation is similar to that of banking and insurance.

This type of mediation exists both in the residential market and in the business market. Interconnect is part of this mediation service where interconnect ensures that all customers in the market can be connected with each other independently of operator.

Simple asymmetric mediation

This case is shown in Figure 9. In this case the mediation is between a 'consumer' market and providers of 'goods'.

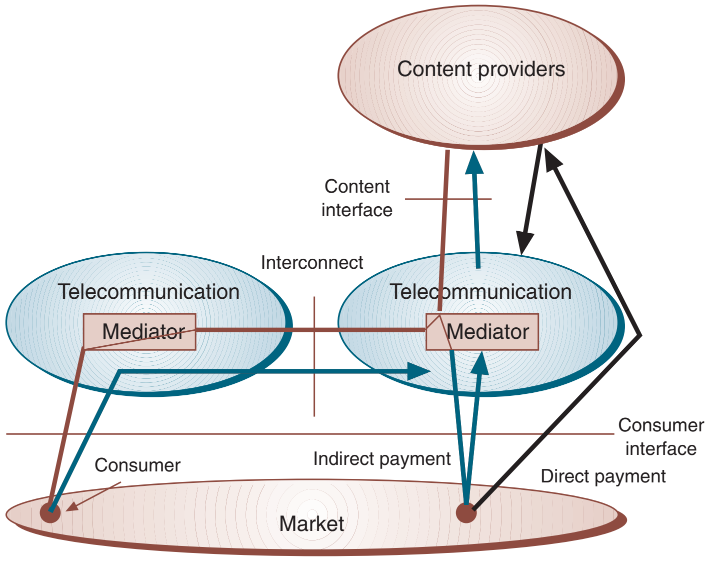


Figure 9 Simple asymmetric mediation

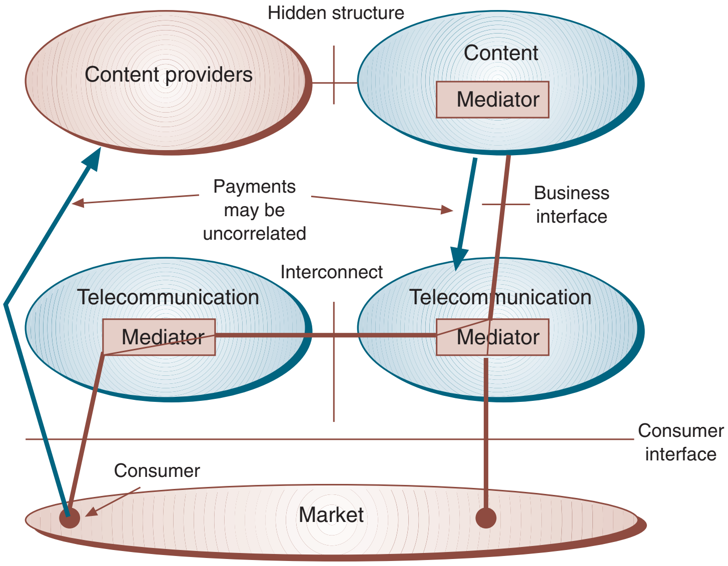


Figure 10 Complex asymmetric mediation

Goods in this respect can be information of any type: video, music and text. What makes it different from broadcasting is that the service is interactive. This is the type of service that exists on the World Wide Web. Premium rate services are also simple examples where you may pay for advise from a lawyer or from an astrologer. It is also the service which is planned for video-on-demand and pay-per-view. The mediation is illustrated in Figure 9. The main point is that the market for the telecommunication operator splits into two markets, each with a different interface: one market toward the consumer and one market toward the content provider. The interface toward the consumer is much the same as in the previous case.

The interface toward the content provider is different and may be so complex that it requires separate system integrators to build it. Payment for the services may be made via the telecommunications operator, or be arranged in such a way that the consumer pays the content provider directly and the content provider pays the telecommunications operator. In this market we may see the first incentives to change the established payment and subscription structures.

Complex asymmetric mediation

This case resembles the former case. The main difference is that the telecommunications operator is one of several mediators interacting in order to deliver the service. The interface toward the content provider is then replaced by an interface toward other mediators. This is illustrated in Figure 10.

The reason why this case is so important is that the telecommunications operator no longer has direct influence on the delivery of content and the way in which telecommunication interacts with the product. Complex products like intelligent buildings, road traffic management systems and co-operative work systems are examples of systems where telecommunications is only one of many components making up the product. Such systems may also require several system integrators in order to establish the complex solution. It will also require co-operation far beyond what is common today.

Note also that payment for the services may be made to some hidden content provider. The telecommunications operator is paid back by the content mediator. This payment may be uncorrelated with the payment from the consumer.

The categories above are related to the interactions between markets and the providers of product elements beyond that of telecommunications. As we saw earlier, the telecommunications companies are in addition split into roles, most of them performing mediating activities: transport provider, access operator, platform operator, subscription manager, contract manager, clearinghouse and content host. The characteristics of all of them are that they deliver parts of the overall product in parallel and at the same time. The same do the other mediators and content providers in Figures 9 and 10. When talking of complex products like intelligent building, co-operative work and electronic trade, we are at the same time discussing a value creating system more complex then any other such system.

In such complex structures there seem to be two main activities:

- Development of the mediating infrastructure;
- Development of services in the mediating infrastructure.

As was mentioned above, these activities may not be done by the value network in the mediating infrastructure but by separate system integrators. Once the infrastructure has been established, there are four generic activities creating the value in the structure:

- *Common project organisation* where the task is to divide the production among the different roles in the value creating network, and to support co-operation between the individual roles, evolution of the common services and markets, and ensure that the quality of the product is retained. In an alliance consisting of energy producers, telecommunications operators, installation companies and manufacturers of equipment delivering remote control of energy to the residential market, this activity may be assigned to a separate company in charge of binding the alliance together and ensuring that the right combination of skills is available for installation, marketing and operation of the system.
- *Development of core value and production management* is the core business of developing a complete product. The activity combines competencies and technologies of all the actors taking part in the production in order to realise the final product. In the above example, this activity may be undertaken by the producers of energy since they will deliver the core element of the product. If the system also delivers a burglary alarm, the activity may be shared among the producers of energy and the companies delivering the safety component.
- *Operation of mediating infrastructure* required for making the product available to the customers. In the example, this activity may be shared among operators delivering communication infrastructure for remote control and operators delivering support for transaction systems and information storage.
- *Customer administration* contains activities like billing and customer care. In the example, this may be handled by several companies specialised in different activities such as billing, maintenance, handling of complaints and promotion. The activity should nevertheless appear as one single activity. This may require an advanced platform for co-operative work.

In complex cases where the product is delivered by an alliance of several companies each accommodating several roles, the basic activities described above may be distributed over several companies. Products may be distributed in different ways, and the same companies or parts of companies may enter several alliances and even be present in competing constellations. Examples of such complex constellations will be found in electronic trade, co-operative work, telemedicine, intelligent buildings, road control systems, and logistic systems. These are all areas hardly developed, where the development may have been hampered by poor understanding of the underlying system of value creation. Note that the examples we are presenting here are several magnitudes more complex than simple mediating industries like banking, though division of this business into specialised roles may create systems of comparable complexity.

This division of responsibility may be found and implemented if an object oriented model is made of the organisation required for making the product. As we saw above, the object oriented model will contain all roles required in making the product and the relationships that must exist between the roles. A sketch of how the activities may be mapped onto the object model is shown in Figure 11. This model only contains a rudimentary description of how a business system for the intelligent building product may be designed. The value production activities may be allocated in different ways; this is just one example. Note also that there may be activities which are not part of the main value creation activities.

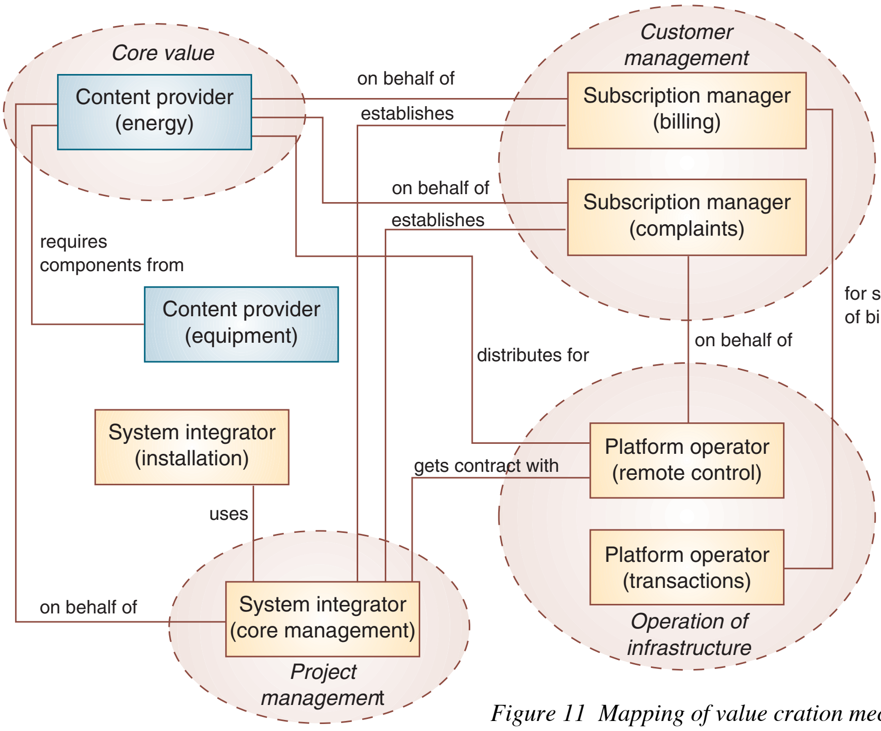

Core value
Content provider (energy)
on behalf of
establishes
Customer management
Subscription manager (billing)
on behalf of
establishes
Subscription manager (complaints)
requires components from
Content provider (equipment)
for support of billing
on behalf of
distributes for
System integrator (installation)
Platform operator (remote control)
gets contract with
uses
Platform operator (transactions)
on behalf of
System integrator (core management)
Operation of infrastructure
Project management

Figure 11 Mapping of value cration mechanisms

6 Conclusions

Above we have only given a glimpse into a new way of looking at mediating industries or – what it is also referred to – value networks. The need for such tools is a result of several things: the increasing complexity of telecommunications, the new way telecommunication interacts with other products, and that some of these products can only be created by alliances combining several skills which have never been combined before.

The use of object orientation is one way of making the complexity comprehensible by only focusing on elements which are core elements of the subject matter, and by suppressing information which are not needed in the given context. Different ways of looking at the problem will require different models, and a complete picture of the system will consist of many models. This way of looking at systems is in fact why object orientation was first proposed in the early 1960s (the Simula language). In our opinion, object orientation is not just one way software can be constructed, but represents a much deeper approach to system design. This approach is now gaining ground since ISO also uses the method for system description, including whole enterprises, in their Open Distributed Processing specification, and not only for design.

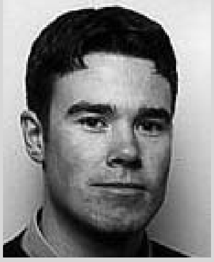
References

1 Audestad, J A. Telecommunications and complexity. *Telektronikk*, 94, (3/4), 1998, 2–20. (This issue.)

2 Rumbaugh, J et al. *Object-oriented modelling and design*. Prentice-Hall, 1991.

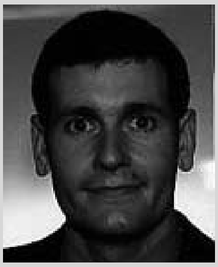
Jan A. Audestad (56) is Senior Advisor for the Corporate Management of Telenor. He is also Adjunct Professor of telematics at the Norwegian University of Science and Technology.

e-mail: Jan-Arild.Audestad@s.hk.telenor.no



Arvid Kjæreng is Business Consultant at Telenor R&D, where he has been working with value creation processes in the ICT industry. His main focus has been to investigate business units' positioning and business roles they can and wish to possess. He has also focused on value creation alterations caused by the technical and content convergence.

e-mail: arvid.kjareng@fou.telenor.no



Laurent Mahieu (28) is Advisor and Project Manager in the Corporate Strategy department of Telenor International AS. He holds a Masters degree in Economics from the University of Aix-en-Provence. In addition, he earned a M.Phil. in Systems Dynamics from the University of Bergen where he conducted research on the dynamics of strategic resources within the telecommunications industry.

e-mail: laurent.mahieu@oslo.international.telenor.no

Real Options and Managerial Flexibility

STEIN-ERIK FLETEN, TROND JØRGENSEN AND STEIN W. WALLACE

The traditional net-present-value (NPV) criterion in capital budgeting has recently been heavily criticised because of its failure to take into account the value of managerial flexibility. In this paper we explain why flexibility is important in real-life investments and discuss advantages and disadvantages of other approaches to capital budgeting, especially focusing on the real-options approach.

1 Introduction

Many investment projects have flexibility that is difficult to capture with traditional discounted cash flow models. The manager may have the option to postpone the start of the project or to alter the project in some way either before or after the project has started. This type of managerial flexibility becomes important for the total value of the investment opportunity when there is uncertainty in the environment of the project. The opportunity to defer a project is the most common example of a real option, where the manager has the right, but not the obligation, to initiate a project at some future date.

Many investment projects can be viewed as real options. Such a view captures the value of flexibility and is particularly suitable for situations where uncertainty is important. In addition real options analysis provides answers to when to use the flexibility and when to wait.

This paper is written for an investment manager who makes capital budgeting decisions for a firm facing an uncertain environment. The manager is willing to formally analyse alternative investment projects and to aid decision-making through mathematical models. She or he is curious about which tools are available for such an analysis, strong and weak sides of these, and is particularly interested in the real options view.

As will be seen in this paper, the real options view is not the only way to consider investment projects. Extensions of discounted cash flow models and decision analysis are related frameworks that can be employed, each having a different focus with different advantages and disadvantages.

The paper is organised as follows: In the next section we demonstrate why flexibility is important, and in Section 3 we explain what real options analysis is and when it can be successfully applied. Section 4 provides an overview and comparison of other tools that can be used in investment situations, and Section 5 provides some examples of investments where managerial flexibility played a role. Finally, in Section 6 we briefly discuss how we can evaluate decisions made in the past.

2 Why managerial flexibility is valuable

By *flexibility* we refer to the ability to make decisions at different stages in time. Generally, decisions depend on what has happened at previous stages. The absence of flexibility means that all decisions must be made initially. Of course, we will generally be better off if we can adapt the decisions during an investment project rather than having to make all decisions initially. The difference in value between these two approaches is called the *value of flexibility* and it can be arbitrarily large as the following example demonstrates:

Suppose we have a risky investment project consisting of two stages such that the second stage can be abandoned if the first stage proves to be a failure. Thus, in this project we have the flexibility of deciding if the second stage will be executed depending on the experience gained in the first stage. Let the profit (net cash flow) of executing the first stage be $c = -5$ and let the profit associated with the second stage be $a = 20$ if the first stage was a success and $b = -40$ if it was a failure. If the second stage is not implemented, the profit associated with the second stage is zero. Of course, the optimal investment strategy in this simple case is to invest in the second stage if the first stage is a success and abandon if the first stage is a failure. If there is a 50 % chance that the first stage is a success, the expected total profit is $(a + 0)/2 + c = 10 - 5 = 5$. So, the project gives a positive expected profit.

On the other hand, in the absence of flexibility, we must decide initially whether both stages will be executed or not. We see that the expected total profit of executing the project now becomes: $(a + b)/2 + c = -10 - 5 = -15$. Thus, the project would not be started in the first place and we get a total profit of zero. The value of flexibility is the difference between the two cases: $(a + 0)/2 + c - 0 = a/2 + c = 5$. We see that the entire value of this project is due to the inherent flexibility! (And we see that this is true for all $a < -b$.)

The above example clearly demonstrates that the value of flexibility should be taken into account in capital budgeting decisions. While obvious in the above example, in more complex situations the value of flexibility is easily neglected. In fact, the classical net present value (NPV) criterion presented in most (if not all) textbooks on capital budgeting is incorrect for projects where flexibility exists! Another way of saying this is that the NPV criterion is only correct in now-or-never investment situations (i.e. where all decisions must be made before the project starts).

A new emerging framework for capital budgeting, replacing the NPV paradigm, models flexibility in terms of options. We discuss such real options in more detail in the next section.

3 Real options analysis

A *financial* call option is a right, but not an obligation, to buy an asset at a fixed price on a future date. If the holder of the option chooses to buy the asset, often termed *the underlying asset*, we say that he *exercises* the option, paying the fixed *exercise price* to receive the underlying asset. This will happen only if the market price of the asset is higher than the exercise price at the *maturity* (exercise) date.

A *real* option is defined similarly, but is more general. The exercise decision is not necessarily to buy an asset, but can be any decision that changes the nature of the cash flows to the holder of the investment opportunity. The irreversible investment cost that is committed at the initiation of the investment project often plays the role of the exercise price, and the underlying asset is often the real project once started. There may be no particular date on which this investment opportunity matures, in which case the investment manager must find the best exercise date.

The most common types of real options are:

- The option to defer;
- The option to abandon;
- The staged investment option;

- Scaling options;
- Switching options;
- Growth options;
- Multiple interacting options.

We will next describe each of these in more detail.

The option to defer the start of a project

Even if a project has a positive net present value (NPV), it may not be optimal to start the project immediately. For example, in a project consisting of constructing and operating an oil platform, the value of the project depends heavily on the oil price. The future oil price is unknown, so the value of waiting can be large. The reason for this lies in the following asymmetry: If we defer the project and tomorrow's oil price is sufficiently higher than today's, then we will start the project. However, if the price drops, we do not need to start the project. We have the option to defer and will wait until (if ever) the oil price is sufficiently high. Generally, the value of this managerial flexibility increases with the uncertainty. Thus, the more uncertain we are about the future, the more reason we have to defer the project.

Example

Consider a fictitious company, TelVal Inc., which has the opportunity to build cable capacity to serve 300,000 long distance telephone calls simultaneously between two cities. This would cost $ 300 million, including the present value of all fixed costs. Suppose the company never has to give up the right to develop this project. The management wants to know how much this project opportunity is worth, and when to initiate it.

There is a spot market for use of such cable capacity, with prices quoted in $/minute. Such a market exists, for example, at www.arbinet.com, and we will assume that these prices obey a certain stochastic process where the logarithm of the spot prices follows a random walk[1]. For simplicity, we assume that there is no variable cost associated with utilizing capacity and that interest rates are constant. The utilization rate is 85 %, meaning that TelVal can sell
60 * 24 * 365 * 0.85 minutes per year on average at the spot rate. TelVal uses a discount rate of 12 % in their capital budgeting.

The method used to solve the project value and optimal development strategy involves dynamic programming and solving a partial differential equation. The details are beyond the scope of this paper, however.

Figure 3.1 shows the result of this valuation.

The dashed line shows the value of the project if the investment decision for TelVal is without the flexibility of deferring the start of the project. The NPV is positive if the spot price is above $ 0.00018/minute. This decision would only occur if the project had a positive NPV, which happens above a spot price of $ 0.00018/minute. The solid line corresponds to the value of the project including the ability to wait until the investment timing is optimal. The difference between these two values is the value of flexibility. The optimal timing of the initiation of the project is when the spot price rises above $ 0.00036/minute (which is twice what the classical NPV theory suggests!).

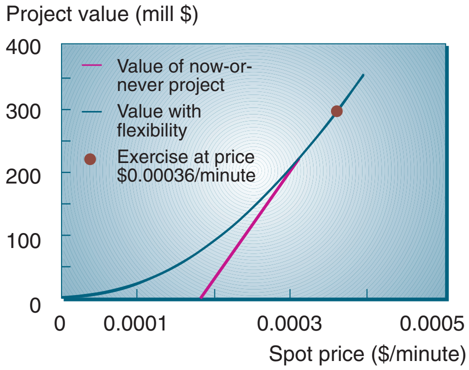


Figure 3.1 Investment strategy and value of cable concession for TelVal's project

The option to abandon a project

If an operating project becomes only marginally profitable (e.g. the net present value is zero or slightly negative), the classical capital budgeting theory suggests we close down (i.e. abandon) the project. However, by real options theory we realise that abandoning a project is a valuable option since it adds flexibility to the project. Since the NPV criterion is invalid for projects with flexibility, it follows that we must reconsider the criteria for abandoning a project. The option to abandon is similar to the option to defer since shutting down a project can be considered as an investment. So, even if the net present value of future operation is (slightly) negative (e.g. operating an oil platform where the costs of pumping up the remaining oil is very expensive) it may not be optimal to close the project down! The reason for this is that the future oil price (on which the value of the project depends) might increase to a level where the project again is profitable. The exact oil price when it is optimal to shut the project down, can be determined using a real options model.

The staged investment option

A common real option in projects with technical uncertainty (e.g. research and development projects) is to divide the project into stages and make a go/no-go decision for the next stage of the project based on the results obtained in the current stage. This introduces flexibility into the project and may enhance the value of the project considerably.

Scaling options

Note that all the real options mentioned so far come free of charge. In other words, these real options are inherent in the projects and no additional investment is necessary to have these options available (although there usually are costs associated with exercising the options). However, in many cases it can be profitable to invest in additional real options. For instance, the option to *alter the operating scale of the project:* For an additional cost, the capacity of a new plant might be chosen at a late stage in the construction process (of the plant) to catch the current trend of the demand. Another important case where scaling options occur is in the general time-cost trade-off problem in project scheduling where the project manager has the option

[1] *The growth rate and volatility are assumed to be 0.04 and 0.2, respectively.*

to allocate more resources to any activity in the project before the activity starts. Thus, the amount of resources allocated to an activity will depend on what has happened prior to the start of the activity (delays etc.) and will be chosen in a way such that the project is completed on schedule.

Switching options

A class of real options very similar to scaling options is switching options. For example, a company might pay the additional cost of planning two mutually exclusive versions of a product such that only one of the plans can be implemented. The choice of which plan to implement may depend on the current demand. Examples of other switching options include the option of switching the technology of production or the channels or targets of marketing the products/services that the real project supply.

Growth options

With the classical capital budgeting theory there is often a contradiction between investment decisions and company strategy. A project might have a negative NPV but still be recommended because of its strategic importance. This paradox disappears when taking a real options approach because it recognises that a project returns not only a cash flow generated by the project itself, but also creates new real options. The new options are often referred to as growth options and allow the owner of the project to execute follow-up projects. For example, a project may give a company an option to enter a new market which it otherwise could not. Such options can be very valuable and may outweigh the costs of the original project. Growth options are ignored by the classical NPV theory and may explain why the market value of some high technology companies (such as Netscape inc.) has been (and probably still is) much higher than what a simple NPV cash flow analysis suggests. Shareholders appreciate the value of growth options even if NPV theory does not.

Multiple interacting options

In a general project, several types of real options are available. Most real option models where the solution is given by a simple mathematical formula consider only a single type of options. However, the value of real options is not additive; i.e. we cannot sum up the total value using many separate simpler models, but need to have one single model integrating the effect of all decisions made in the project. Such complex models usually require extensive numerical calculations, which in practice are only possible by computers.

4 Alternative tools

To maximise the value for shareholders, the investment projects with the highest value should be selected. All decisions, both in selecting projects and in carrying them out, should reflect this criterion. Thus the investment and project managers should be making optimal decisions under uncertainty. In this situation it is natural to consider using *stochastic programming* techniques, a field which focuses on optimisation and mathematical modelling of problems having uncertain parameters.

Advantages of stochastic programming include modelling flexibility, simultaneous valuation and optimisation of the project at hand and use of forecast scenarios. Disadvantages include the fact that no clue is offered on which discount rates are to be used. A fixed rate is often used.

Using a constant cost of capital, or discount rate, is also common in the *discounted cash flow* (DSF) method of investment analysis. In this method, one finds the expected cash flows through the life of the project, discounts them to present values, and adds them up.

By varying the assumptions of which information is available, different stochastic programming models can be formulated. Capital budgeting models can be divided into three groups, depending on their underlying assumptions:

1. Dynamic discounted cash flow (DCF) models;
2. Decision analysis (DA) models;
3. Real option valuation models, also known as contingent claims analysis (CCA) models.

All these models, if used correctly, can take into account the value of flexibility. However, note that none of these models correspond to the traditional approach of estimating a single cash flow and then applying the net-present-value (NPV) criterion. We refer to the traditional models as NPV models.

The DCF-based models assume the existence of several different cash-flow scenarios, e.g. represented as a 'decision tree'. However, there are two major problems with DCF models. First, the probability that a certain scenario will occur must be specified explicitly (for all scenarios). Estimating the probabilities is usually difficult. Second, DCF models depend on estimating risk-adjusted discount rates. Under certain assumptions, this can be done using the Capital Asset Pricing Model (CAPM), but it is very complicated in practical situations consisting of a large number of different scenarios. Simplifying the analysis by using only a single discount rate can lead to large errors.

DA models are similar to DCF models in that both models can use 'decision-trees' for specifying cash-flow scenarios. Thus, the problem of estimating the probabilities remains. The difference between DCF models and DA models is the way risk is handled. DA models use the risk-free rate of return as the discount rate. This captures the time-value of money, but not the uncertainty. To model the effect of uncertainty DA models introduces so-called *utility functions* and show how a single decision-maker should act to maximise the expected utility of the profit. This approach is based on certain axioms for rational behaviour of a single decision-maker. However, a problem with this approach is that it is not market-based but rather based on the subjective utility function of a particular decision-maker. In situations where a single decision-maker cannot be identified or in situations where we want to maximise the market value of a company, we are outside the scope of DA models.

CCA models (i.e. real option models) overcome the problem of a single subjective decision maker (in DA models) by assuming the existence of underlying tradable assets (e.g. commodities). Contingent claims analysis (CCA) is based on analyzing a risk-free portfolio obtained by purchasing one unit of the project and selling a sufficient number of units of the underlying asset such that the variance (risk) of the portfolio is zero. Based on this principle and using a stochastic model describing how the price of the underlying asset evolves over time, CCA models can determine, among other things, the conditions when the project should be started (as opposed to being postponed). The optimal condition for when to start a project usually takes the form of starting the project as soon as the price of the underlying asset (e.g. commodity price) reaches a certain level. In order to employ contingent claims analysis, the risk in the underlying project has to have a market price. This is because the risk (or uncertainty) of the option depends on the risk in the underlying project. In fact, the option derives its value only because there is uncertainty in the underlying project. The risk[2] is priced, for example, if the project is to produce a storable commodity sold in a perfect market, and the only important source of project uncertainty is the future price of the commodity. If the product or service is not storable, there may still exist a futures or options market having the price of this product as the underlying reference, in which case the price of risk is reflected in the prices of futures or options. A practical problem with CCA models is obtaining a realistic stochastic model for the price dynamics and the price of risk. This can be done by analyzing historical data via statistical methods. Choosing an arbitrary stochastic model can lead to arbitrary results. Thus, back-of-envelope calculations using simple formulas obtained by using a particular stochastic model, are dangerous[3]. Another practical problem with CCA models is obtaining the rate of foregone earnings. For financial options this corresponds to the dividends paid.

In all the three models – DCF, DA and CCA – stochastic programming techniques give the tools needed to obtain the optimal decisions under uncertainty. Stochastic dynamic programming is a good example of such a tool. In addition, advanced statistical methods are used for describing the dynamics of prices. A lot of effort has been made to develop analytical solutions to capital budgeting problems for important special cases. However, a numerical approach (using a computer) is necessary for more general problems.

5 Some real-life examples

Flexibility is generally important in all situations where we are making decisions under uncertainty. We have already discussed different methods that can be used to deal with such flexibility (including real options valuation models). In this section we present four, very different, real-life cases demonstrating the existence of flexibility: 1) in copper mining; 2) in the construction of a power plant; 3) in the design of offshore oil platforms; and 4) in the cold war!

- **Canadian copper mines** have been extensively studied using real options theory [4]. Copper mining is a risky activity due to the very volatile copper price. However, if the copper price gets too low, the management has the flexibility to temporary shut down the mine. Later, when the copper price is again sufficiently high, the mine can reopen. This *option of temporary closure* is neglected by traditional NPV models.
- **The Svartisen Power Plant** was developed by Statkraft from 1987 to 1997. It is based on hydro-electric power and has a generation capacity of 350 MW, which means that it must run for about 7000 hours on average each year to avoid reservoir spill. The original plans were to have twice this capacity, 700 MW, to go with the large reservoir capacity. However, Statkraft used the flexibility they had available to adjust the generation capacity after initiating the project and after seeing the changes in the profitability (but before the end of the development). In other words, they *exercised a scaling option*. At the same time, they decided to buy the flexibility to reverse this decision by reserving space in the power station house and building tunnels for installing a second 350 MW generator at a later date. In real options terminology: They purchased a *growth option*. Although real options theory was not used, the decisions were certainly subject to formal analyses at Statkraft.
- **Design of offshore oil platforms** must be done under the uncertainty of the nature of the reservoir. There can be more sand or water in the reservoir than expected, the size of the field is unknown, and future oil prices are uncertain. Thus it may be profitable, although very expensive, to reserve some blank space on such a platform in case extra equipment is needed later for removing sand or water, or for producing at a higher rate.
- **The cold war** between the United States and the former Soviet Union during the decades following the Second World War is perhaps the most gigantic example where flexibility was essential. Both sides purchased the option to annihilate the other by installing intercontinental ballistic missiles (ICBMs) armed with nuclear warheads. From a military point of view, having this option had a large positive value because it would prevent the other side from attacking. In this case, formal real options theory is not applicable since it is not a market based problem. However, real options theory gives us a useful terminology for describing the situation. A more suitable formal approach in this case is applying *game theory* (i.e. an extension of decision analysis that deals with problems where the outcome is affected by *several* decision-makers as well as uncertainty).

6 Evaluating decisions made in the past

Both in the power-plant example and the cold-war example the options purchased were not exercised (to date). Looking back in time and evaluating decisions made, it may now be obvious that significant costs of purchasing such and similar options could have been saved *if we had known that the options would never have been exercised anyway*. By making this argument, it may seem tempting to conclude that the options should not have been purchased. However, making such a conclusion would not be rational! At the time the decisions were made, the decision-makers *did not know* if they would exercise the options or not, due to the *presence of uncertainty*. In retrospect, the only way we can find out if an *optimal* decision was made is by analysing the situation using only the information available to the decision-maker at that point in time. The scenario that actually did occur, was only one of several possible at the time the decision was made. It was, of course, beyond the decision-maker's control to choose the scenario! If the information available for the decision-maker is not available in a later evaluation of the decision, we cannot know if the decision was optimal or not, even if we know what happened later! Thus, learning from history is not trivial.

[2] *Or more precisely, the part of the risk that is impossible to diversify away through other investments.*

[3] *This conclusion is supported by analysis regarding the profitability of Canadian copper mines [4].*

7 Conclusion

In an uncertain environment *flexibility* in investment projects becomes more valuable, and for managers to capture these values and make better decisions it is necessary to employ models that incorporate both uncertainty and sequential decision-making. Because investments have a strategic importance for firms, those firms that are able to utilise and acquire flexibility wisely will have a better chances of surviving in the long run.

References

1 Christiansen, D S, Wallace, S W. Option theory and modeling under uncertainty. *Annals of Operations Research,* 82, 1998. (In preparation.)

2 Dixit, A K, Pindyck, R S. *Investment under uncertainty.* Princeton, NJ, Princeton University Press, 1994.

3 Kall, P, Wallace, S W. *Stochastic programming.* Chichester, John Wiley, 1994.

4 Slade, M E. *Managing projects flexibly : an application of real-options theory.* Vancouver, Univ. of British Columbia, Dep. of Economics, 1998. (Discussion paper No. 98-02.)

5 Trigeorgis, L (ed.). *Real options : managerial flexibility and strategy in resource allocation.* Cambridge, MA, The MIT Press, 1996.

Stein-Erik Fleten (27) is a graduate engineer from NTNU 1995. He is a resarch fellow at the Department of Industrial Economics and Technology Management at NTNU and is working on a Ph.D. thesis with the working title "Theory of portfolio management in the electricity industry".

e-mail: sef@iot.ntnu.no

Trond Jørgensen (28) is a doctoral student in Operations Research at NTNU. He holds a Master's degree from the University of Oslo, 1994. He has worked 3 years at Norwegian Defense Research Establishment in the Satellite Remote Sensing Group. He also has experience in unsupervised classification in multispectral satellite images. Other jobs include computer programming and teacher in computer programming at high-school level. His main interest is stochastic dynamic decision problems arising in complex systems.
e-mail: tj@iot.ntnu.no

Stein W. Wallace (42) has been Prof. of Operations Research at NTNU since 1990 after spending 7 years as Senior Scientist at Chr. Michelsen Institute, Bergen, and 3 years at Haugesund Maritime College. Throughout, his focus has been on theoretical and applied aspects of decision making under uncertainty, with applications in fisheries management, petroleum field development, portfolio management and telecommunications.

e-mail: sww@iot.ntnu.no

New Aspects of Service Provision and Technology Strategies in Telecommunications

SHANE DYE, ASGEIR TOMASGARD AND STEIN W. WALLACE

Because of deregulation and new technology the telecommunications markets are changing. As the computer industry and the telecommunications industry become more similar, services based on processing of information become more important than before. This means that there is a need for new technology strategies and new decision support tools at both the operational and the strategic level. In this paper we describe a new approach for decision support at the operational level of service provision when services are processing based. We particularly treat the case where demand for services is uncertain, discussing how the flexibility of new distributed processing environments can be used to deal with this. Strategic effects of such an approach are briefly discussed. An example model is given to illustrate resource allocation in a distributed processing environment.

1 Introduction

New possibilities for providing services in distributed telecommunications networks have appeared as a consequence of technological developments such as: digital technology, modern packet switched high speed networks, architectures like ATM [4, 16] and standardization of software and equipment. See [13] for a discussion on the technological aspects of distributed processing in networks. While [13] does not focus on telecommunications, the Telecommunications Information Networking Architecture Consortium (TINA-C) [2, 17] is an example of an initiative defining a future distributed telecommunications standard.

The new technological aspects have not only introduced new possibilities in telecommunications but, along with deregulation and free competition, they have contributed to a change in the environment where the telecommunications industry finds itself. Competition has become harder; new players enter the scene and the roles of old players change.

Traditionally, switching and bandwidth are considered important bottlenecks. Currently, transportation capacities are being increased through the use of fibre optics, advanced protocols and data compression technology. With the parallel emergence of new services such as video on demand, pay per view television and video conferencing, it is highly unlikely that the future will hold telecommunication networks without transportation bottlenecks. However, many new services and some transportation technologies require more and more processing resources, thereby creating a new bottleneck. Add to this the computing industry's influence upon the telecommunication market, and you get both additional computational power and competition. Together, this makes it difficult to predict which of the bottlenecks will prove to be most influential in the future.

There is much ongoing research investigating ways to increase transportation capacities and make efficient use of the transportation resources available. However, while there is a push to provide increased processing resources there is little research into the efficient use of available processing resources. This is the issue we address. In particular, we review the situation where computational resources provide the only bottleneck.

When considering processing resources there is little telecommunications literature. In [12] the distribution of the workload in a network of computers is discussed with whole jobs being processed on a single machine. In [18, 19] modelling of distributed processing in service provision is examined. Related to the models given there, an algorithm to find a feasible service configuration for the case where demand is known and all demand for services must be met is given in [8]. Approximation algorithms for the case when demand is known and profit is to be maximized are given in [5]. The only known approximations for the case when demand is uncertain is when services are installed only at one node [6]. As we will see later in Subsection 4.3 this single node case has specific and important interpretations also in a distributed processing setting.

This paper is based on the results in [5, 6, 18, 19]. Section 2 defines the technological framework we operate within. Then we look at service provision in Section 3. The main focus is on an operational approach for service provision trying to utilize the flexibility of a distributed processing environment. The strategic effects of such an approach are then discussed. In Section 4 we examine a model to illustrate the service provision approach and give some results on the hardness of such models.

2 Distributed processing in telecommunications

Aspects of distribution are crucial to the modelling in later sections. Here, we give a brief description of the distributed framework used for the discussion and the models of later sections.

2.1 A distributed network model

The relationship between the transportation network, the computing nodes with processing resources and the applications used to build services can be represented by Figure 1.

The model consists of three planes. In the upper plane we have the set of interacting applications which form an interconnected structure. Links here are representative of the interactions between the applications in order to provide services. In

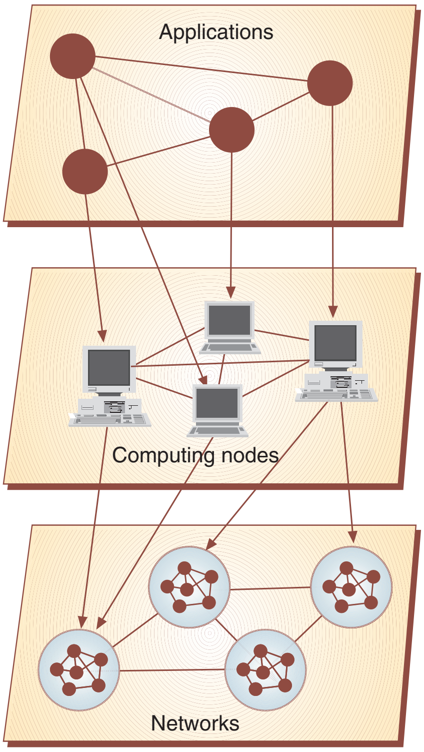


Figure 1 The relationship between the transportation network, the computing nodes and service applications

the bottom plane we have a set of interconnected networks describing the physical connectivity between the applications. The computing nodes, where the applications reside, can be seen as the 'glue' which binds the distributed application architecture and the network architecture together. The (all-to-all) links in this plane show which computers have access to each other through the underlying networks.

Note that the mappings between the three planes may be many-to-many mappings, i.e. one computing node may accommodate several applications and several identical copies of an application may reside at different nodes. Similarly, one computing node may be connected to several networks, and one network may accommodate several nodes. Furthermore, these mappings can be dynamic. Applications may be installed or removed from computing nodes. Also, systems are emerging where even the mapping between nodes and networks can be dynamic; for example, in the form of computing nodes residing in mobile terminals or low orbit satellites.

It is worth noting that this distributed network model provides a useful model for the Internet and World Wide Web applications. The upper plane corresponds to the Web, the middle plane represents the servers, and the bottom plane is the international structure of telephone and data networks.

2.2 Distribution implications

The exact technicalities of the distributed environment are not important for the modelling discussion. Instead, we base our models on a framework implied by the distribution model. The important aspects are as follows.

- Applications which together constitute a service can be placed on different computing nodes without affecting the interaction between them.
- Applications may be dynamically relocated, even whilst providing processing and interaction with other applications.
- Several copies of an application may be present in the network.
- The processing of a single customer (or customer group) may be dynamically moved from one copy of an application to another without affecting the customer's service or other interacting applications.

Such properties will be implied by many distribution frameworks. In particular, they are implied by the *transparencies* of TINA-C [1, 3, 15].

A further assumption about the distribution that we make use of in various models is

- The processing of an application may be provided externally without affecting the customer's service or other interacting applications.

Applications which are provided 'externally' might be provided by a different part of the service provider's network or by a different service provider.

2.3 Distribution of services

We define a *subservice* as a collection of applications that will always be located together to perform their assigned task. In the rest of this paper we will consider the subservices as the smallest indivisible entities of computer code that are the building blocks for the construction of services.

Therefore, all services are built from subservices, as shown in Figure 2. Here service A is built from subservices 'sound' and 'video'. Service B is built from subservices 'sound', 'data' and 'video'. The distribution transparencies mean that services can be distributed, with their subservice instances spread over several nodes. Figure 2 illustrates this for service B. Also several instances of the same subservice may exist at different nodes, as for subservice 'sound' in the illustrated example. The arcs in the lower half of the figure represent the different instances. A subservice instance can be used by several services and a service usually consists of several subservices.

3 Service provision, network design and distributed processing

Firstly, we discuss more specifically the task of allocating network resources to services in order to meet demand. We start with describing two types of resources – node resources and transportation resources – before we examine a resource allocation approach utilizing the network's distributed processing capabilities. Thereafter, we study how resource allocation at the operational level influences network design.

3.1 Service provision and limited resources

A service is built up by interacting subservices. The subservices and their interaction require both transportation and computational resources. Typical transportation resources are bandwidth, switches, servers, transmission options

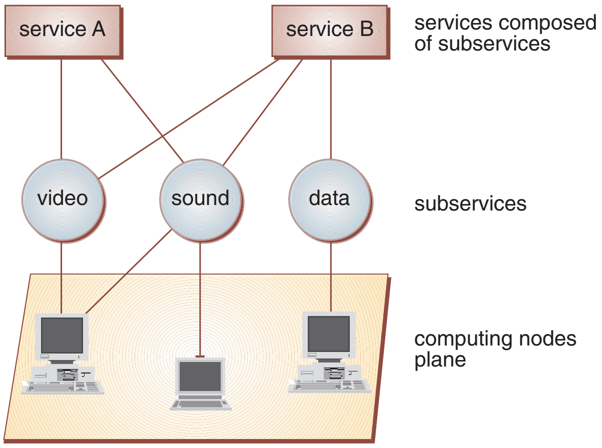


Figure 2 Services are composed of subservices which are distributed in the network

and routers. Computational resources are provided at the network's computing nodes; examples include processing capacity, memory and input/output capacity.

The customer's requirements from a service are covered in the concept *Quality of Service* (QoS) [10]. QoS for applications or subservices can be looked at as a set of user perceivable attributes defining the desired behaviour of the subservice when it comes to, for example, delay, reliability, image quality, image size, security etc. At a lower level one also defines QoS parameters for the networks: bandwidth, delay, blocking probabilities, loss rates, etc. In light of the properties of Section 2.2 we assume that the QoS of a subservice is not affected by the specific computing node on which it is installed.

The properties from Section 2.2 indicate that the location of the subservice instances used to meet requests is not important from the point of view of the interacting subservices. Additionally, customers have no preferences when it comes to the service instances they use, except through QoS. The distribution therefore gives rise to a great deal of flexibility when it comes to allocating resources in the network, and are crucial to the problem described in this section.

We make the fundamental assumption that we have a distributed network in which the properties from Section 2.2 are valid. If processing capacity is the only limiting resource, the following assumption makes sense:

A1: In regions of the network, a service's demand for processing capacity can be served at any computing node in the region, independent of the demand's and the node's locations and without affecting the service's QoS.

This in turn implies the following underlying assumption:

A2: *Within* regions of the global network, the transportation networks are designed to keep the time delay acceptable on transportation of generated information and to ensure the correct QoS for services, independent of the allocation of requests for subservices to nodes.

When computing capacity is not the only limiting resource, these assumptions provide a conceptual framework for decomposing a global network into smaller subnetworks for which the following models are valid.

It is also interesting to note that the models may remain useful even without assuming that computing resources are the bottleneck. This is because the network provider is not necessarily the service provider. A service provider may be free to rent computing capacity from the network provider. In the same way that the location of service processing may be hidden from the customer, the transportation structure may be hidden from the service provider. Then the service provider's limiting resource is solely the purchased computing capacity.

The main conclusion we can draw from the above discussion is that given the assumptions, customer demand for processing capacity can be met at any computing node independent of the location of demand.

3.2 Network design and critical resources in a region

The flexibility of the above approach at the operational level also has strategic effects. The choice as to which resources limit a telecommunication system is not arbitrary. The network's design is vital to providing the balance between network resources. Here, we provide a brief discussion of network design. We focus on the degree of distribution of the processing capacity at the strategic level and its effect on flexibility at the operational level.

The discussion here is to highlight the trade-off between transportation and processing from the point-of-view of network design. Given this perspective, we do not include existing network resources in the models. In practice, the current infrastructure must be considered and it brings to the models additional restrictions relating to such things as mixing new and old technology, keeping services available while improvements are being made and legal aspects restricting or requiring change.

Finding an 'optimal' network design is not straightforward. The trade-off between transportation and processing resources is complex. Primarily, there is the choice of how much to invest in each of these resources. Then, remember that within a region like the ones defined in Assumption *A2* of Section 3.1 the customer is indifferent to where demand is met. This allows for 'trading' of processing capacity between different locations of the region: processing capacity installed at one location can be used to meet the demand from several other locations. This is useful when different parts of the region experience peaks at different times of day, like for example a city in business hours and its suburbs after work. In addition trading gains benefit from the economies of scale and utilizing resources in low investment cost areas. If demand is uncertain and not completely correlated in the different parts of a region, trading of capacity is one efficient way of dealing with the uncertainty. Capacity trading of course increases the need for transportation resources, as information must be moved before it is processed. Even if trading of capacity in a distributed processing environment means more flexibility and is one of the most important tools to deal with uncertainty both at the operational and the strategic level, the extra overhead in transportation must be considered.

Additionally, network designs which tend to use the transportation resources most efficiently often make processing a bottleneck, and vice versa. This is illustrated in the following extreme examples.

Since the network is distributed by nature, control information and signalling are reduced as a consequence of a distributed architecture. This reduces the transportation load. An extreme possibility, therefore, is to use only local processing with many small computing nodes. The effect is little transportation, but low utilization of processing capacity as each subservice will be installed in many locations. Typically there is an overhead in terms of resource use connected to running a subservice, for example a fixed use of capacity independent of the number of customers. This overhead will now be induced everywhere. At the other end of the spectrum, using one enormous, centralized computing node only requires one copy of each subservice. However, signalling increases and all services must send information through the centralized node, both of which create a bottleneck in transportation resources. In addition, the centralization raises reliability issues.

We try to describe trading of processing capacity and the trade-off between processing and transportation capacity using an hierarchical approach to network design by considering two models. The 'local' model determines investment in network infrastructure (telephone lines, switches and computing nodes) within a (small) region of the network. Typically, this would be a region as defined in *A2*. This requires the infrastructure to allow the chosen demand to be met with a given level of QoS. The increased transportation infrastructure cost from trading processing capacity within the region is weighted against savings in installing computer capacity.

The 'global' model determines how to invest in network resources and allocate demand among these various regions. We also allow trading of processing resources between regions, again at the expense of transportation resources. Clearly, in this model we implicitly assume transportation between regions to have a limited effect on QoS. This is valid as the transportation links are planned for meeting the extra demand for transportation capacity that is induced by trading. These models are detailed in Tomasgard et al [18, 19].

For the remainder of this paper we implicitly assume that the investment problem is well solved and that we are free to utilize the computational resources in the network as efficiently as possible through the use of distributed processing.

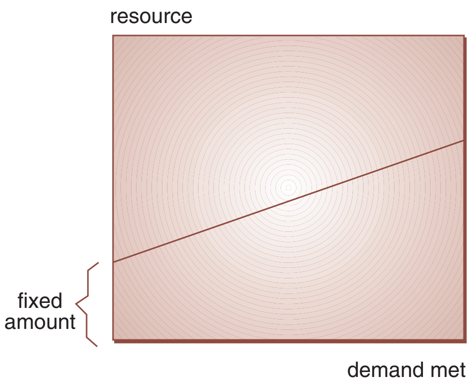


Figure 3 A subservice's use of processing capacity with increasing demand met

4 A service provision model

We will now present a modelling approach to service provision of processing based services as described in Section 3.1. Afterwards we discuss uncertainty, how to deal with it, and some results on computations. This section is based on [5, 6, 19].

4.1 Assumptions

To provide a clear and consistent model we make the following assumptions. The number of requests for a service describes the demand from the customer locations. Demand for a service induces a corresponding demand for the subservices used by the service. The demand induced for a subservice differs from service to service. In the model we only implicitly consider the service demands through their combined demand for subservices. We allow subservice demands to be met either internally or externally. A demand met internally is processed by a subservice at one of the computing nodes. A demand met externally does not use any node resources. Instead, there is a single, linear cost associated with the externally met demand. In practice, this demand will either be met in another region of our network or we will pay another service provider to meet this demand for us. In the latter case a spot-market exists for trading subservices. One can imagine several types of competition. We will here assume that none of the service providers are able to influence the price by their decision, and regard the market price as fixed. If a more complicated price structure is assumed, the models given here will be a crude approximation.

When a subservice is installed on a node, the only limit on the number of requests that can be served by it arises from the node's processing resources. In respect to this, there is *one* constraining processing resource, referred to as processing capacity. We distinguish between two types of processing capacity used by a subservice. The first is a fixed amount (called the fixed resource requirement) induced by having the software running and ready to meet requests. A subservice uses this capacity whenever it is installed on the node, even when it does not satisfy a single request. The second type of capacity use comes from accessing the subservice. This requirement increases in direct proportion with the subservice's demand met at the node. The combined capacity use of a subservice is illustrated in Figure 3.

4.2 A deterministic model for several computing nodes

We first describe a static model assuming deterministic demand. The main reason for considering such a simplistic model is that it gives us insight into the underlying allocation problem by being easier to analyse. These insights will provide useful guidance for more realistic but more complicated models where the demand fluctuates over time and is not known in advance.

The model we present is a linear mixed integer programming problem (MIP) [14]. The aim of the model is to best allocate the subservices to the computing nodes. Subservice demands which are not met by the computing nodes are met externally and we define the best allocation as having the least cost for demand met externally. There are m computing nodes and n subservices. The zero-one variable z_{ij} is 1 if subservice j is installed on computing node i and 0 if not. The variable x_{ij} denotes the amount of demand for subservice j met at node i, while t_j is the demand for subservice j met externally. Let the demand for processing capacity for subservice j be d_j. The cost of processing subservice j externally is q_j. The fixed resource use by subservice j is denoted by r_j. The total capacity of node i is s_i. M is an arbitrarily large number which provides an upper bound for all x_{ij}. We formulate the model as follows:

$$min \sum_{j=1}^{n} q_j t_j$$

$$\text{s.t.} \sum_{j=1}^{n} r_j z_{ij} + \sum_{j=1}^{n} x_{ij} \leq s_i, \;\; i = 1, ..., m,$$

$$t_j + \sum_{i=1}^{m} x_{ij} = d_j, \quad j = 1, ..., n,$$

$$M z_{ij} - x_{ij} \geq 0, \quad i = 1, ..., m, \;\; j = 1, ..., n,$$

$$z_{ij} \in \{0, 1\}, x_{ij} \geq 0, \quad i = 1, ..., m, \;\; j = 1, ..., n$$

$$t_j \geq 0 \quad \bar{v} = 1, ..., n \qquad (1)$$

The objective function calculates the total cost of meeting demand externally. The first set of constraints ensures that the computing node capacities are adhered to for each node, while the second set of constraints requires that demand is either met externally or internally at some computing node. The third set of constraints forces nodes which meet the demand for a subservice to have that subservice installed.

4.3 Different interpretations of the model

In this model we implicitly assume that the service providers have full knowledge of the sizes of all the computing nodes and the ability to utilize them as they see fit. This may be the case, for example, if the service provider and network provider are one and the same. When the service provider is not also the network provider, the situation differs depending on the level of control allowed to the service provider. The service provider may just hire a certain guaranteed quantity of processing capacity from the network provider. The service provider can decide what subservices to provide and levels of demand to meet, however, the network provider is free to replicate and migrate the subservice instances running on his infrastructure. The service provider sees the infrastructure only as a single node. The above model is still valid, but now $m = 1$. Single node models are discussed in [5, 6, 18, 19].

However, the network provider must still solve a multi-node problem. By hiding the complete infrastructure from the service provider the network provider is completely free to best utilize his resources, possibly combining the needs of many service providers. The cost to the network provider of such a scheme is that the implicit costs of subservice replication are not directly borne by the service provider. His gain is that he is free to reuse subservices for several service providers.

4.4 Modelling uncertainty

Understanding the effects of uncertainty on a system is an important step to fully understanding the system.

One of the main advantages offered by the distribution transparencies of Section 2.2 is the possibility of dynamically changing the resource allocation. This ability to change is even more important when we try to consider uncertainty of demand in our models. In the treatment of uncertainty here and in the rest of the paper, we use terminology from stochastic programming [11]. In a stochastic programming setting we assume that uncertainty in the demand for services can be modelled as random parameters with known probability distributions.

At any time the *state* of the system of computers running services can be described by the current demand for processing capacity, a probability distribution for future demand and the current whereabouts of subservices. In this paper we focus on the case of discrete distributions; each realization of demand is called a scenario [11] and can be interpreted as a demand vector with an assigned probability to occur.

When a decision is made to install a subservice at a node, the subservice cannot usually be used to meet the demand immediately. There is a set-up time in which the subservice must be collected from a database, loaded into the memory and initialized, before it can be used to serve requests. There can also be a delay between the time the service provider decides to remove a subservice and the time its fixed resource use is released. Under specific assumptions [19] concerning set-up times, shut-down times and the uncertainty structure, the decision process can be modelled as a two-stage process. In the first stage we decide which subservices to install on which computing nodes, having only probabilistic information about demand. During the subservices' installation time uncertainty is resolved. At the second stage we must meet the realized demand using the subservice configuration chosen in the first stage. This decision process is illustrated in Figure 4. In the figure, at the point where demand is revealed, only one of the two scenarios S1 and S2 will be realized. We cannot know beforehand which path will occur.

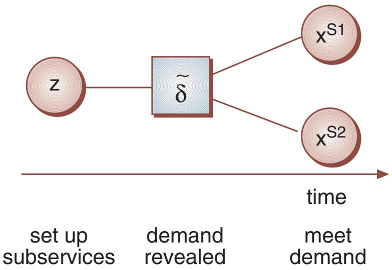


Figure 4 An example of the two-stage decision process. Here there are two possible scenarios, S1 and S2

The current value of a given first stage decision is found by evaluating the cost of meeting demand in the second stage for each modelled scenario in the best possible way, *given the fixed first stage decision*. For each first stage decision, we see a probability distribution of second stage costs. The optimal first stage decision has the least expected second stage cost. For a further introduction to modelling of dynamic decision processes using scenarios see for example [20].

4.5 Heuristics and computational complexity of the models

In the previous section we advocated the incorporation of uncertainty in the model without indicating how one would solve the resulting model. In fact, adding uncertainty to a model generally makes it more difficult to solve. The single node deterministic problem ($m = 1$) can be solved efficiently by dynamic programming [5]. The algorithm's running time is pseudo-polynomial, that is the number of steps is bounded by a polynomial in the number of subservices and the node's size. However, the multiple node deterministic problem is hard to solve to optimality. Typically, it requires a strictly exponential method such as branch and bound. The stochastic multiple node problem is at least as difficult as the deterministic. Decomposition methods are used in [9] to accelerate the solution process. For the stochastic single node case no pseudo-polynomial method is known, but decomposition based on efficient algorithms for subproblems may even here be a promising approach to solve the problem efficiently [7]. If we consider the number of scenarios as fixed and not as a problem input, even this problem can be solved in pseudo-polynomial time.

It is clear that currently there is no hope of obtaining an algorithm that solves the service provision problem to optimality in a real-time decision system. Indeed, theory suggests that there will always be a limit to the size of problem that can be solved in reasonable time. Consider that currently a deterministic single node model with around 250 subservices will take up to a few seconds on a Sun Ultra 2 machine running at 200 MHz. Similar sized deterministic multiple node problems or single node problems with stochastic demand have less predictable solution times because of the methods used, but they typically vary between seconds and several hours.

A real-time system would need to use heuristics or rules of thumb as decision support. These provide fast but non-optimal solutions. The trade-off is between quality and speed of solution. In practice, even so called 'optimal solutions' are not always the best, as models never precisely portray reality. So the trade-off between speed and quality is not always possible to accurately estimate.

One method for evaluating the effectiveness of a heuristic is to consider their worst-case performance ratio, a measure on how badly they can perform. The heuristic solution is measured against the optimal solution. To compare heuristics, a bound on the running time is also used. This allows users one way to evaluate the trade-off between speed and quality. Good heuristics or approximation methods give solutions that are no more than a known constant multiple of the optimal solution, with running times bounded by a lower order polynomial.

We now discuss some previously studied heuristics for the service provision problem from [5, 6]. For this discussion we consider a slightly changed problem. Instead of minimizing the cost of external servicing, the profit from internal servicing is maximized. Here the profit for meeting one unit of demand for subservice j internally is q_j, the cost of buying the same servicing externally. That is, we consider what we have saved ourselves by not purchasing all processing externally.

First we examine a greedy heuristic for the deterministic case. Defining

$$\bar{q}_j = \frac{q_j}{1 + r_j/d_j},$$

this can be viewed as a measure of profit per capacity unit used if demand is met completely for a subservice. The greedy heuristic for the single node case is simple. The subservices are sorted by decreasing $\bar{q}_j$ and the node is filled in this order until fully meeting the demand before the next (critical) subservice violates the node's capacity. The heuristic then chooses whether to install all subservices up to the critical one or simply install the critical one alone. This gives a solution that is at least half the optimal solution. The maximum running time is proportional to the number of subservices. For the multiple node problem the nodes are considered in any order. The critical subservice for the current node is left out and the next node is filled. The heuristic chooses whether to install this solution or one of the critical subservices singly. The greedy solution is at least $1 / (m + 1)$ times the optimal solution. The maximum running time is proportional to the sum of the number of subservices and the number of computing nodes.

For the stochastic single node problem, a similar heuristic is based on solving the linear programming relaxation of the problem. The solution time is still polynomial. Here we choose to install all subservices for which $z_j = 1$ or any of the subservices where $0 < z_j < 1$ at the optimal solution of the LP relaxation. This leads to the heuristic being at least $1 / (k + 1)$ times the optimal solution where k is the number of scenarios. More advanced heuristics with a constant performance ratio exist even for this problem, where the heuristic solution can still be calculated in polynomial time.

For the deterministic single node service provision problem it is possible to approximate in polynomial time the solution value within $1 - \varepsilon$ of the optimal solution value, utilizing dynamic programming. Here ε can be chosen to be as small as desirable, but the solution time of the approximation algorithm increases in $1/\varepsilon$. For the stochastic problem theoretical results show that there is little hope of finding approximations that always give near optimal solutions [6].

Current work suggests that it is still possible to find good heuristics for the stochastic problem. These heuristics can give us knowledge which is possible to implement as rules of thumb in a real-time system. Finding such heuristics and rules of thumb is the focus of future work.

5 Conclusions

We identified the need for optimization models in telecommunications as a consequence of deregulation, new competition and new technology. Because of a change in focus from transportation to processing for many new services, such models should include distributed processing oriented aspects of a telecommunications network. One reason for this is that the computer industry and the telecommunications industry become more and more similar.

Given the assumption that processing capacity is the limiting resource in our network, we conclude that, in regions of the network, allocation of computer resources to services at an operational level may be made independent of the location of the demand. We suggest how the increased flexibility coming from distributed processing and trading of capacity can be used to dynamically allocate resources in a distributed network.

Increased flexibility clearly influences the decisions to be made for investments in infrastructure, both when it comes to processing and transportation capacities. Strategic decision models for distributed telecommunication networks cannot be studied isolated from models concerning dynamic resource allocation and the assumptions they are based on. A shift in focus at the operational level from transportation of data to processing of information is also bound to present a need for modelling new aspects when it comes to investments.

References

1 Audestad, J A. *Lecture notes : Course 45356 Communication in distributed systems.* Trondheim, UNIT/NTH, Department of Computer Systems and Telematics, 10.10.1995.

2 Barr, W J, Boyd, T, Inoue, Y. The TINA initiative. *IEEE Communications Magazine,* March, 1993, 70–76.

3 Chapman, N, Montesi, S. *Overall concepts and principles of TINA.* TINA-C, February 1995. (Document no. TB_MDC.018_1.0_94.)

4 De Prycker, M. *Asynchronous transfer mode : solution for broadband ISDN.* NJ, Prentice Hall, 1995.

5 Dye, S, Stougie, L, Tomasgard, A. *Approximation algorithms and relaxations for a service provision problem on a telecommunication network.* Trondheim, Norwegian university of

science and technology, Department of industrial economics and technology management, 1998. (Working paper #2-98.)

6 Dye, S, Stougie, L, Tomasgard, A. *The stochastic single node service provision problem.* Trondheim, Norwegian university of science and technology, Department of industrial economics and technology management, 1998. (Working paper #3-98.)

7 Dye, S, Stougie, L, Tomasgard, A. *Single node service provision with fixed charges.* Trondheim, Norwegian university of science and technology, Department of industrial economics and technology management, 1998. (Working paper #4-98.)

8 Dye, S, Tomasgard, A, Wallace, S W. Feasibility in transportation networks with supply eating arcs. *Networks,* 31, 1998, 165–176.

9 Dye, S, Tomasgard, A, Wallace, S W. *The stochastic service provision problem : Benders decomposition.* 1998. (In preparation.)

10 Halsall, F. *Data communications, computer networks and open systems.* New York, Addison-Wesley, 1992.

11 Kall, P, Wallace, S W. *Stochastic programming.* Chichester, Wiley, 1994.

12 Kreidi, A, Sansó, B. *Optimization of a geographically distributed air-ground airline telecommunication system.* Technical report, GERAD publication G-94-47, 1994.

13 Mullender, S (ed.) *Distributed systems.* New York, Addison-Wesley, 1993.

14 Nemhauser, G L, Wolsey, L A. *Integer and combinatorial optimization.* New York, Wiley, 1988. (Wiley-Interscience series in discrete mathematics and optimization.)

15 Nilsson, G, Dupuy, F, Chapman, M. An overview of the telecommunications information networking architecture. In: *TINA 95 Conference, Melbourne,* P1, 1995.

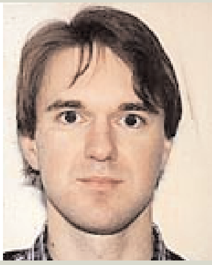
16 Onvural, R. *Asynchronous transfer mode networks : performance issues.* Boston, Artech House, 1994.

17 Rowbotham, T. The TINA Consortium : a collaborative way forward. In: *TELECOM 95,* Geneva, 1995.

18 Tomasgard, A et al. *Stochastic optimization models for distributed communication networks.* Trondheim, Norwegian university of science and technology, Department of industrial economics and technology management, 1997. (Working paper #3-97.)

19 Tomasgard, A et al. Modelling aspects of distributed processing in telecommunication networks. *Annals of Operations Research,* 82, 1998, 161–184.

20 Wallace, S W. The role of uncertainty in strategic planning. *Telektronikk,* 94, (3), 1998, 21–23. (This issue.)



Shane Dye (29) graduated as B.Sc. in mathematics from Canterbury University, New Zealand in 1991 and finished his doctoral thesis at Massey University, New Zealand in 1994. After holding a Post Doctorate at the Dept. of industrial economics and technology management at the Norwegian University of Science and Technology from 1995 to 1997, he now holds a position at the Dept. of Management, University of Canterbury.

e-mail:

Asgeir Tomasgard (28) graduated from NTNU in 1993 and is currently studying for his Ph.D. at the Dept. of industrial economics and technology management at NTNU with a thesis entitled "Aspects of service provision and distributed processing in a telecommunications network". This year he has been working part-time at SINTEF Industrial Management, Economics and Logistics, where he will start as full-time researcher in December 1998.

e-mail: at@uit.ntnu.no

Stein W. Wallace (42) has been Prof. of Operations Research at NTNU since 1990 after spending 7 years as Senior Scientist at Chr. Michelsen Institute, Bergen, and 3 years at Haugesund Maritime College. Throughout, his focus has been on theoretical and applied aspects of decision making under uncertainty, with applications in fisheries management, petroleum field development, portfolio management and telecommunications.

e-mail: sww@iot.ntnu.no

How to Survive in the Future

JAN A. AUDESTAD

1 Introduction

A company lives on its portfolio of products. If its portfolio is attractive to the market, the company may live well and be successful. However, societies change and the market with them. The current portfolio may then be without value in a few years.

One key element in evolution of companies is to prepare the road to the future. This is certainly not simple in telecommunications as we saw in [1] – the problem being the rapid and complex evolution of the area. This complexity exists in all parts of the industry: technology, market, competition, organisation and interaction with society.

This paper is mainly based on two external sources: the definition and analysis of competencies [2]; and ways in which to manage innovation [3]. Important material is also taken from [4], [5] and [6]. Some of these ideas were used in Telenor to determine a strategy for how to assess both current assets of the company and which competencies are required to navigate safely into the future.

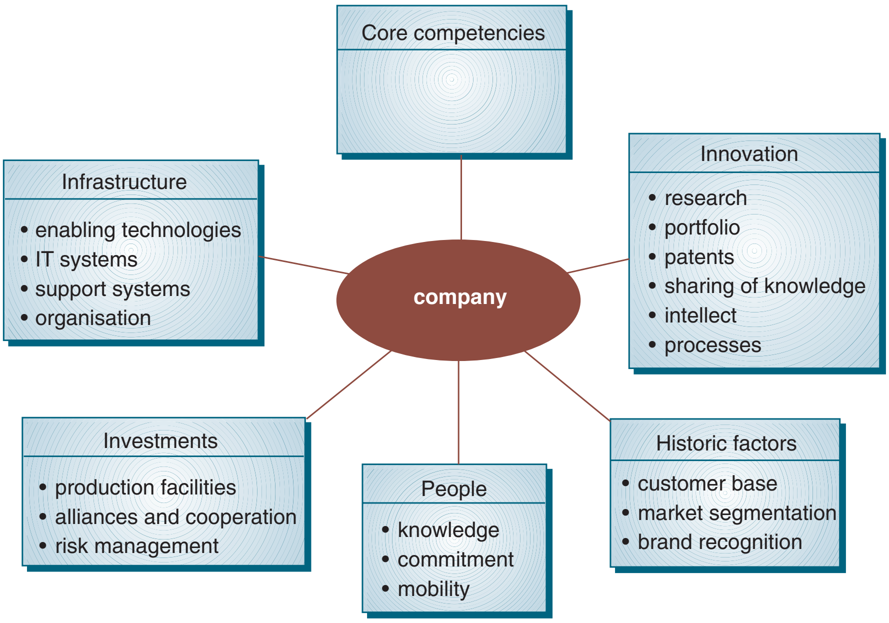


Figure 1 Assets

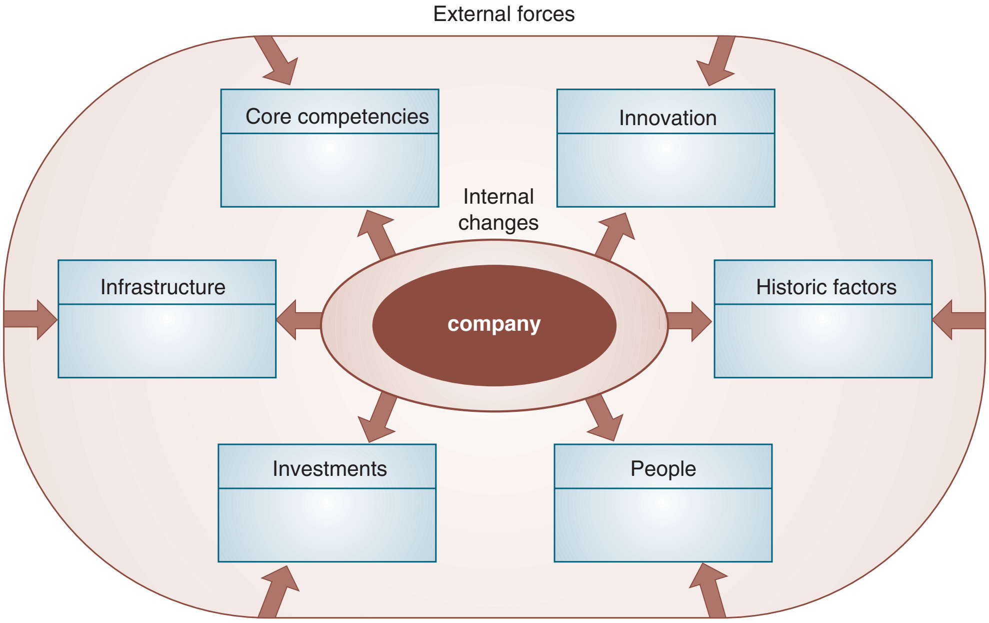


Figure 2 The company is not a stable unit

2 Categories of Assets

A company may possess assets as shown in Figure 1.

The six categories are

- Innovations, manifested by research, portfolio of products, patents, sharing of knowledge, intellect and processes;
- Historic factors such as customer base, market segmentation and brand recognition;
- People defined by their knowledge, commitment to the company and their mobility;
- Investments defined in broad terms as investments in production facilities, alliances and co-operation with similar or other companies, and risk management;
- Infrastructure consisting of technologies required to produce the products, IT systems, support systems and the organisation of the company and its different parts;
- Core competencies which are the bundle of technologies, skills and intellect that will ensure success in the future.

In order to determine the strategy the company needs to consider all these elements. It is likewise important to assess continuously how well these elements are mastered and whether the set is being changed as a result of external forces or by evolution of the company itself. This is illustrated in Figure 2. External forces and internal evolution are not independent since internal changes may be a result of external forces.

The forces shaping the company are thus not static but change with time. This makes the management of these forces so difficult.

Attempts at innovating are made continuously in a company. The innovation is there to increase and renew the portfolio of products, improve the processes of manufacturing, promotion and sale, and differentiate the company and make it unique in the marketplace. The bigger and more rapid the changes in the field are, the more innovative the company must be in order not to lag behind and lose the competition. This means that companies in rapidly changing industries, such as telecommunications, must keep a high level of research and keep a staff of high intellect. By intellect we mean here the capacity to create and use knowledge [3]; it is not mere skills required to produce and manage. In a rapidly developing industry the way in which knowledge is distributed and shared is becoming very important for success. Knowledge must be shared in order to reach a critical mass where it becomes useful. Knowledge is shared by education, co-operation, conferences, publishing, networks of acquaintances and so on. This is particularly difficult in telecommunications because of the combination of new and immature competition and rapid evolution. This has set back joint evolution by several years and it will take time before more fruitful co-operation within the global telecommunications industry may be re-established. In a round table debate of ITS98 it was pointed out that much money is being wasted on research programmes resulting at best in small innovations with minor economic importance. This seems to be a world-wide phenomenon, and it is assumed that a company can take leadership in the field only if it is able to lift its research from short term to real long term.

Some knowledge is legally protected by patents though this protection is usually short term in the telecommunications industry because the products protected may be of value only for a limited time; the patents are in many cases easy to circumvent often resulting in better products, and they may mask the need for innovations and set back the evolution of new, competitive products.

What we do now soon becomes history. Every mistake or success of the present may subtract from or add to the image of the company and shape the brand recognition of the future. The image of the company formed at present is therefore probably the most important factor shaping the future. The history of the company is a function of time in the sense that every future value is not only determined by the value of today, but also by earlier values. In some cases the historic influence may be short; in other cases it may be long. For example, the way we segment the market today is a function of how we did this historically and it is changed only according to the response fed back by the market; in this case the change may depend much on historic factors. On the other hand, the change in the customer base is less dependent on history (and in this way is more similar to an initial value problem in differential equation) but more on what happens just now: increase in competition, decline or improvement in customer care, marketing of replacement products, and so on. In mathematics, it is much easier to understand and solve initial value problems than problems where the solution depends on a string of historic events. The same can probably be said for strategic problems: they are easier to handle if the outcome only depends on what we do now and not on what we did yesterday and the day before.

One example illustrating the complexity of temporal evolution is given in [5] where the causes of the downfall of Encyclopaedia Britannica since 1990 are used as a good example. Historically the Britannica had a large customer base, but had not understood how the customer base was built up and segmented, and which external forces had formed it. When these forces changed – because of the availability of cheap information on the Internet – much of the customer base collapsed since one of the important variables in the 'customer base function' was that parents did not buy the Britannica because of its intellectual value but to do the right thing for their children. The variable 'do the right thing' changed from buying an intellectual product to buying computers with Internet connections where apparently equivalent intellectual products were available together with many other 'right things'.

Organisations need knowledgeable people; that is, people who cannot only perform the tasks the organisation requires in order to produce the products and sell them in the market, but also develop the organisation into the future. In [3] knowledge is divided into five areas of increasing importance:

1 Cognitive knowledge, or the 'know what': the rules and facts of a discipline;

2 Advanced skills, or the 'know how': the ability to perform a task sufficiently well;

3 System understanding, or the 'know why': the understanding of the relationships between key variables;

4 Motivated creativity, or the 'care why': the capacity to interrelate disciplines and create new effects;

5 Synthesis and trained intuition, or the 'perceive how and why': the capability to understand or predict relationships that are not directly measurable (see also [7] where this aspect is addressed in depth).

One example of this ladder of knowledge is information technology. The 'know what' can be how to use Microsoft Office products to create texts and drawings; the 'know how' can be the skill to install, manage and maintain such systems in a corporate network; the 'know why' can be knowledge and understanding of how the corporate network interacts with the Internet and is part of the Web; the 'care why' can be the understanding of how this system is used for electronic trade, telecommuting and co-operative work; the 'perceive how and why' may be how to implement such a system over many actors for jointly delivering services to the intelligent building: comfort, security, payment, encyclopaedia, dictionary services, pay-per-view television, banking, and so on.

In organisations producing advanced and complex products, the importance of the knowledge increases at least exponentially down the list; that is, towards motivate creativity and synthesis. On the other hand, it is exponentially more expensive to educate people in order to give them the right advanced skills and cognitive knowledge (the 'know what' and 'know how' end of the scale). It is, therefore, cheap to create an organisation with much intellect but expensive to train people to do the day-to-day tasks. Therefore, it is common to neglect the higher forms of knowledge and focus the company on developing the concrete skills. Such organisations may be competitive now but not in the future because they lack the intellect required for change. We see this often in job advertisements. Most companies want people with specific skills to do specified tasks (mastering of database X, programming language Y and operating system Z). What happens when these skills have to be replaced by new skills because the portfolio has changed? It is very rare that a company simply asks for clever people! It is also well known that talent attracts talent: if an industry is the best in its field it is easy for them to get the best people. The cost and value in investing in people are also studied in [8]. The findings are that intellect is cheap; that skills are often valued too high and awarded so as to create wrong incentives; and that the companies with much intellectual skills are among those rated highest in the stock market.

Ideally, a company should invest in products that give the highest revenue. This is, of course, not possible and the company has to select a set of products giving a reasonable income over time. This also means that the portfolio has to change over time. In some industries, like telecommunications, the change is rapid and difficult to assess; in others it is slow and predictable. In order to operate in a rapidly changing environment the company may need to form alliances with other companies in the same sector or create new businesses by forming alliances with companies in other sectors. The business will change over time, and so will the alliances; alliances will sometimes also be formed with competitors and non-co-operative games may be replaced by co-operative games, or a mixture of co-operative and non-co-operative games. In such environments, risk management will be a key factor. Probably, new methods need to be taken into use since the risks and parameters determining them are so complex to assess. As we saw in [1], such extreme complexity is indeed present in telecommunications, and this complexity is new so that little can be learned from history. I have not found any other industry in modern time changing so rapidly over such a short time as telecommunications.

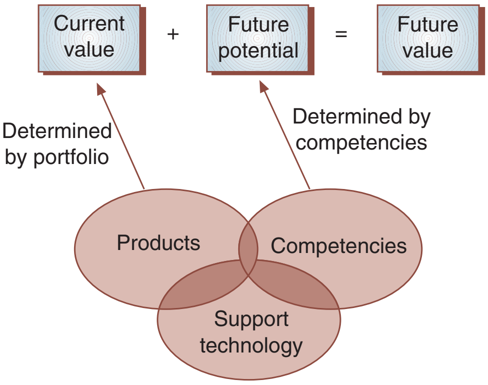


Figure 3 The future value

Much of the success of modern organisations depends on the infrastructure of the organisation. This infrastructure consists of technologies enabling the production of the goods, IT systems and other support systems, and the way in which the company is organised. Telecommunications industries are facing a special problem in this respect: much of the enabling and support technologies are the products created by these technologies; that is, the products made by the industry are the enabling technologies required to produce the same product. Some examples are: decentralisation of a telecommunications company requires telecommunications products marketed as co-operative work and telecommuting; the telecommunications operator needs a business communication network, or intranet, in the same way as other organisations; it uses its own products such as networks to create other products called services; and it builds into these products IT components used as support systems for creating these products. This is different from other companies producing support systems: a company producing 3D drawing and analysis software sells these products to industries using them to produce cars, ships and aircraft. On the other hand, oil companies use their own gasoline for their cars distributing this fluid; building constructors use buildings they have built themselves. However, here we are talking of products that are much simpler than telecommunications products.

This leads to an unusually high degree of internal trade and co-operation between business units in a telecommunications company. The co-operation is also qualitatively difficult to manage for the reasons just mentioned. This management is not made simpler since the telecommunications company is not a Porterian chain but rather a value network [4], favouring different ways of co-operating internally and externally.

The final force shaping the company is the core competencies. We will soon come back to these.

3 Methodology for Determining a Company's Assets

One way of assessing the value of a company is as shown in Figure 3 [2].

The formula given in Figure 3 is simple. It states that:

1 The current value of a company is directly given by its portfolio of products;

2 The future value of a company depends on the portfolio and the way it is developed into the future (the future potential). The future potential is determined by the competencies.

In the formula we have included the support technologies which may be different for different products and competencies. This is in line with [3].

The current value is high if the portfolio has high market penetration and creates a reasonable margin. The future value is high if the company has the competencies required to keep a high position in the market also in the future. In other words, the position in the market is high in the future only if the competencies are such that the future portfolio will become the leading one in the marketplace.

The current value may be valid over a period of time where there is reasonably little uncertainty in the estimates of the market. In telecommunications this time may be several years in certain market segments and for certain products (telephony in the residential market), but very short (one year or less) in other segments (for example, segments related to the Internet). The changes we have seen recently in the telecommunications market indicate that the time where we may assess the current value to be reasonably stable is about one to two years.

On the other hand, the future potential must be estimated over a long time in order to identify the really important competencies. Note that one such competency may just be the ability to make assessments in an environment with much uncertainty. Reasonable planning horizons for the future potential may then be five or more years: The horizon should not be too close because then the changes may not be big enough. To identify some critical competency, the horizon should not be too remote because then important competencies may be blurred by excessive uncertainty.

We have also included the support technology in the formula. This element was not included in [2] but its importance is indeed recognised in [3]. The support technologies are required both for creating the product portfolio and for supporting the competencies. IT systems are the basis for products like electronic trade and electronic newspapers, and telecommunications technologies are required for distribution and remote control of such products. Competencies like customer care, sales management, logistics and customer base management all rely on IT systems.

4 Current Value Analysis

The current value of a company may be characterised by two types of diagrams: one diagram relating the margin and the growth potential of all the products of the company; and one diagram showing price and differentiation potential for each product. The first diagram is shown in Figure 4.

The growth parameter may be the estimated growth rate at present. The estimate may be based on the average taken over some years (the compound annual growth rate). The other axis shows the margin. Each product may be visualised as disks with a size proportional to the overall revenue of that product in the market. The market share of the company under consideration and its competitors may be shown as sectors.

Figure 4 shows four products. Product 1 has a high margin but low (or even negative) growth potential. This is a typical cash cow that should be exploited now, but small investments should be allocated in this area since the product is likely to disappear. Product 2 has a low or negative margin or shows little or negative growth. Such products do not contribute to the wealth of the company, neither now nor in the future, and should be terminated. The problem with such a product is that the termination cost may be high, or that the company is obliged to keep it for fulfilment of licensing requirements.

Product 3 offers high margins and much growth potential, and is thus attractive from a business aspect. These are products that the company should protect or capture. Product 4 has small margins and much growth. This product is then likely to become a problem for the company in the future unless its profitability is improved. The strategy for such products is to change their profitability or, if that is not possible, withdraw them from the market.

The products in this diagram are not static. Over time they may move from one category to another; for example, a product of type 3 may become a cash cow or become unprofitable. This dynamics must be built into the model in order to prepare the company changes in the portfolio.

The vulnerability may be assessed by use of a price differentiation diagram as shown in Figure 5.

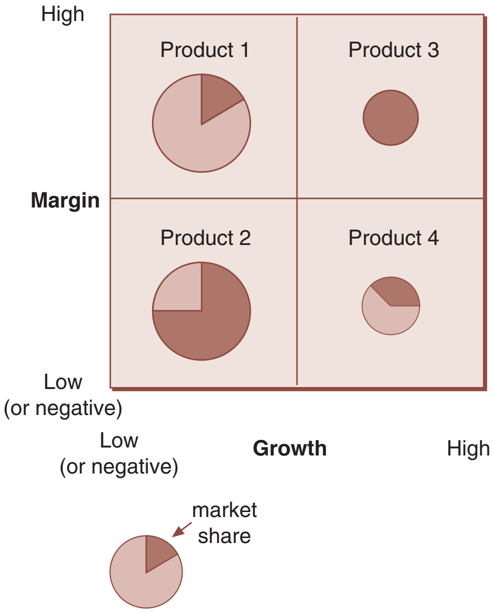


Figure 4 Margin vs. growth

Functionality may replace price and vice versa. This is illustrated in the diagram by splitting it into two regions, the blue region and the red region. In the blue region the company in question has advantage either because the price is low and the product is well differentiated (product 1); the price is so low that it compensates for the superior functionality of the competitor (product 2); the product is superior from a differentiation point of view and the price is not so high that it masks the differentiation advantage (product 3). For a product in the red

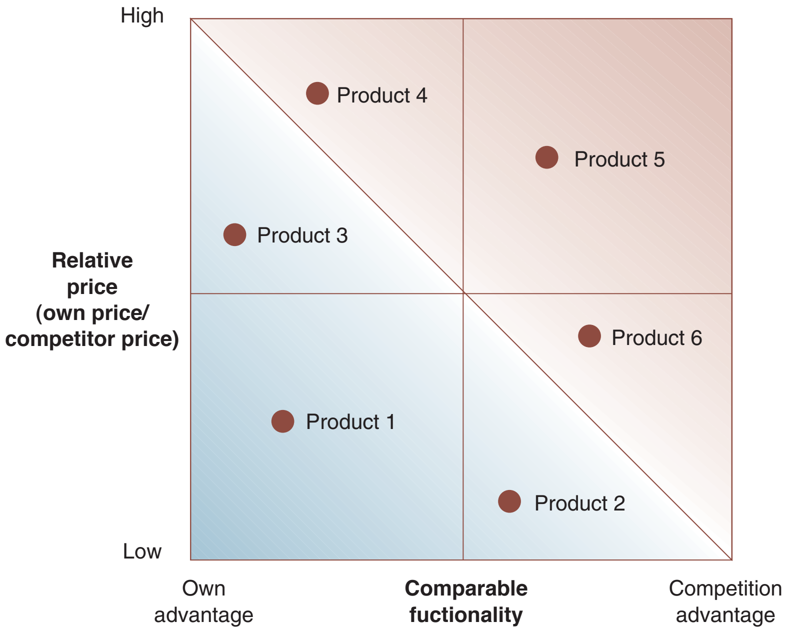


Figure 5 Vulnerability

area the following applies: the price is high and the differentiation advantage is not sufficient to compensate for it (product 4); and the product has neither enough differentiation advantage nor price advantage (products 5 and 6).

5 Future Value: Competencies

The competencies are the collection of skills and knowledge that enable the company to succeed in the market. One systematic way of determining the competencies is a series of investigations of the following items:

- The vision of the company;
- The key success factors;
- The industry drivers;
- Customers and customer interfaces;
- The benefits provided to the customers;
- Identification of competencies;
- Consolidating and ranking the competencies;
- Determining the company's mastering of the competencies;
- Propose how holes in the competencies can be filled;
- Build up enthusiasm for change in the organisation;
- Introduce changes as required in the organisation.

Background information may be obtained by use of workshops where executives and key personnel of the company participate to create a common view of the company, the industry sector to which it belongs, and the forces that will shape the industry. The process is then as follows.

First, the long range goals of the company are defined and compared with the vision formulated for the company. In this process it may be found that even the vision should be adjusted, for example, in order to meet new expectations of the market. To achieve the vision the company needs to see or measure which factors are determining the success of the company. These are the key success factors. It is important that these factors can be measured in order to assess how well the company is able to fulfil the vision. These factors may be:

- Financial measures;
- Customer satisfaction and market share;
- Efficiency of operational processes;
- Organisational learning and assessment of the knowledge base of the company.

It is important that other measures than financial are considered because it will normally be found that the core competencies are related to the mastering of these measures ([2], [3]) rather than the traditional financial measures. As is easily observed in the market, the financial status of a company may change so fast that it is impossible to build a long term strategy upon it.

The complex dynamics of telecommunications is driven by a large set of parameters called the industry drivers. These drivers are external and one reason for defining them is that the company needs to define responses to these drivers or adapt to them as required. The drivers are then the same as we defined as external forces in Figure 2. The responses are the internal changes.

The key drivers may be grouped in aggregated sets as for example:

- Market demand variables comprising the impact of national or international economic conditions, integration with products of other industries, identification of supplementary, complementary or replacement products, and demographic changes;
- Technological variables such as speed of introduction of new technologies, technical agility of competitors, investment requirements and capital binding, and the impact of unexpected technological evolution;
- Regulatory variables including pricing constraints, co-operation requirements, and delivery obligations in non-profitable markets.

The impact of the changes is related to and may be understood in terms of catastrophe theory. Catastrophe theory was proposed by the French mathematician Thom during the 1950s. The theory cannot easily be used to predict a given behaviour but it may help understand qualitatively certain types of sudden changes which may occur. The mathematics and several applications of catastrophe theory are found in [9] and [10].

The ideas behind the theory are:

1. Almost all systems will develop along the most natural path and almost all shapes in nature will take the most natural form from a mathematical point of view;
2. A generic or canonical description exists for such forms;
3. Many such forms are structurally stable; that is, they are not altered by small changes;
4. This generic view also applies to the properties of the space in which systems develop.

Catastrophes are classified into six families of elementary forms (Thom's classification theorem). Non-elementary and more complex forms exist, but they may not be structurally stable. To us, the most interesting of these families is the two-parameter cusp catastrophe, since several of the analyses we have done in Telenor indicate that the future is determined by two aggregated, free variables (for example, market growth and regulation) and one dependent variable (margin). Though there are many detailed variables, they tend to aggregate into a few variables describing the system in enough detail. Similar behaviour is seen in the development of fish stocks [9], where the external variables are the efficiency of the fishing method and the number of fishing vessels. If we take, for instance, the performance of the company as the dependent variable and the two parameters (or free variables) market growth and regulation as independent variables, the performance of the company as a function of time can be viewed as a trajectory on a two-dimensional surface as shown in Figure 6.

According to Thom's classification theorem the most natural form of this type of surface will contain a cusp singularity as shown in Figure 6 (the argument for this conclusion in too long and complex to be reproduced here). The behaviour of the company may be as indicated by the two trajectories. In some cases the evolution will be smooth (trajectory 1). If the trajectory passes over the fold (trajectory 2), there will be a sudden change in performance. What is characteristic for this case is that the changes may be small for a large range of variation of the two parameters (in the smooth areas to the right or left of the fold) so that it is difficult to foresee the catastrophe from merely observing the performance. If the trajectory is reversed, the performance will change as shown by the dotted line. The existence of the hystheresies makes the performance irreversible. This is exactly the behaviour observed for the fisheries in the North Atlantic Ocean. It may also explain why a company may prosper one day and be bankrupt the next day.

However, as was said above, this is only a qualitative description of the behaviour. It will probably be impossible to both find the catastrophe surface and the trajectory describing the performance of the company. The importance lies in that this description will make management and others aware of this type of non-linear behaviour and build it into the strategy plans.

The company needs to define their customers accurately. In the telecommunications industry a vertical company like Telenor will have several types of customers requiring different management:

- Residential customers who may be divided further into segments requiring different handling;
- Business customers requiring telecommunications as a primary product;
- Business customers using telecommunications as production factors in their own primary products (electronic commerce, bank transactions and distance learning);
- Other network operators requiring interconnect;
- Service providers requiring infrastructure services in order to complete their product.

For each customer group the company should analyse in depth what the group requires and how well this requirement is met.

From these analyses it may now be possible to identify which competencies the company requires in order to progress safely into the future. In this way a considerable set of competencies may be defined. The competencies may then be ranked as shown in Figure 7.

The example is not taken from any existing business or company; it is included only for visualisation of the ranking. The competencies may be ranked along two axes: the abscissa indicates how important the competency is for enabling leadership and the ordinate is assessing

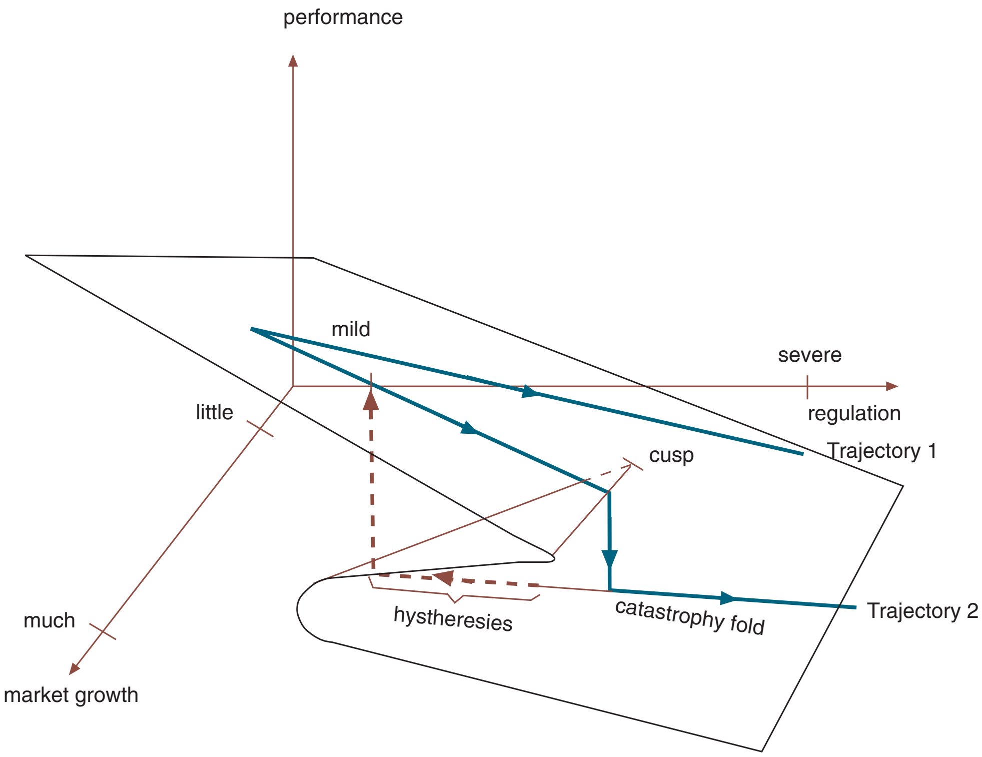


Figure 6 Performance trajectories

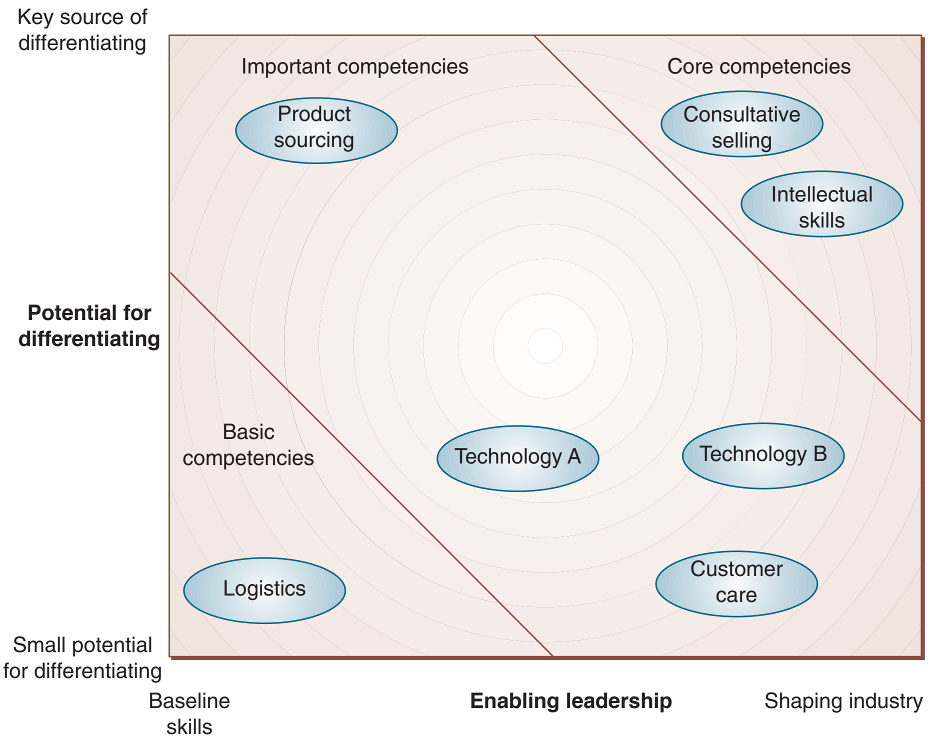


Figure 7 Classification of competencies

its potential for differentiation. Some competencies are just baseline skills (logistics in the example) which the company must possess but has no impact on how the industry will evolve in the future. On the other hand, there may be competencies (such as technology B and intellectual skills) which will shape the future. Between these extremes there may be other competencies which may be more than basic skills but with only limited impact on the future evolution.

On the ordinate we show the competency's potential for differentiating one company from another. In the example, logistics is regarded to have a small impact on the potential for differentiation. Product sourcing and consultative selling are, on the other hand, competencies which are key sources of differentiation.

The diagram may be divided into three parts as shown: basic competencies, important competencies and core competencies. This division is important in order to identify where the company should focus its strategy. As part of this analysis the company should assess how well the different competencies are mastered in order to identify where resources should be placed for improved competition and 'world mastery'. Building basic and important competencies will be normal educational activities in the company. Building core competencies will be a strategic endeavour.

Competencies can be obtained in different ways. Some competencies may be home-grown; that is, developed by the company itself by funding research and educating employees. Competencies can also be obtained by alliances where competencies the company lacks are obtained from the partners. There are also cases where a competency is present in part of the company. In such cases the competency may be spread to other parts of the company by leveraging the competency. Finally, competencies may be bought or hired by buying a company (for example in logistics) or hiring individuals or consultants possessing the competency.

6 Characteristics of Core Competencies

It is the core competencies that will make the company successful in the future. They are the main factors that will ensure endurance of a company's leadership in the market.

We saw that the core competencies are those competencies which differentiate the company from its competitors and is the reason why its products may be preferred by the customers. At the same time, it is the competencies which will form the future of the industry.

In addition, the core competencies are difficult to copy by competitors either because they are expensive to create, difficult to build up, or difficult to understand.

Successful companies target few core competencies – often only two or three, and seldom more than five. The core competencies are not the traditional functions such as finance, sales, production or engineering, though these are the activities around which organisations were formed in the past [3]. Instead, they are multifunctional activities cutting across the organisation not favouring power-building by product or management groups trying to gain dominance of the company. In this way, the core competencies are not products but rather intellectual competencies that can sustain a leading edge even if some products wither and disappear or are replaced by others. Such competencies may then be based on skills in technologies, methods, processes, market assessment and so on, where each such competency does not support only one but several products.

One such core competency may, for example, be the management of cannibalising own products [11]. Such cannibalisation may be the most important incentive for radical innovations. It is difficult not because it creates new businesses but because it may remove part of the business. In an industry where the changes are fast, the understanding and management of cannibalisation may be the most important core competency. Hewlett Packard is one example of a company where this competency is well developed.

An example of the utilisation of technical competency and marketing competency is Honda [3]. Honda is a leading company in small engine design and innovative marketing and distribution. These competencies combined provide the company with a core competency which has been difficult for other companies to match.

The core competencies must be such that they result in values to customers in the long run. It is argued that one core competency of every company should be related to understanding or serving the customers. Several computer firms with the world's best technology in its niche, have failed in just this area because they have either had difficulties serving the market (Cray Research) or adapting to changes in the market (Norsk Data).

The value of core competencies may become even more important when an industry is disaggregated. This has happened in the biotechnology industry by coalition of several specialised companies in order to create an easily manoeuvrable aggregate industry, where each element may develop along their own core competencies. This has also been the case for a long time for the watch industry in Switzerland and the furniture industry in Norway. Here there are companies specialised in making parts and being so good at doing this that no other companies can match them on neither quality nor price. Now the same evolution is coming in telecommunications where the regulation of the liberalised market has first of all prepared the ground for companies operating in narrow and profitable niches [1]. Telecommunications, being an industry of value networks, is a good candidate for decomposition of the industry into niches which may lead to new types of co-operative behaviour in the market. The development is still too new to assess the result of such decomposition.

The new development called C4 convergence, where the four Cs stand for Content, Computer, Communication and Consumer electronics, may give rise to much more co-operation and competition. One core competency may then be to discover, establish and maintain co-operation in such markets. This core competency may be supported by the competencies of bundling products and technologies, binding customers through part of the portfolio contained in the bundle, stimulating buyers by letting some part of the offer being free of charge (as for some Internet products), and create a shared market recognition which is not associated with a single product but with the synergy obtained by the combination.

7 Conclusion

The main concerns of a company in the telecommunications market should now be:

- Understand the highly dynamic changes now taking place in technologies, markets, external regulations and competition, and recognise that survival in the field will require excellence in strategic awareness and manoeuvrability;
- Increase the intellectual value of the company by recognising that the cost of the company is linked to the detailed 'know what' and 'know how' skills of the company, while the future success depends on how well the company is organised around intellect (the 'care why' and 'perceive how and why' knowledge) noting that it is cheap from a financial viewpoint but difficult from a management viewpoint to retain the best intellects of the organisation;
- Keep an accurate overview of the portfolio of products with regard to sustainability, margin created by each product and vulnerability to competition. This is a dynamic process requiring regular analysis in order to detect changes which may require strategic decisions;
- Identify and create core competencies which can ensure a safe journey into the future.

References

1 Audestad, J A. Telecommunications and complexity. *Telektronikk,* 94, (3/4), 1998, 2–20. (This issue.)

2 Prahalad, C K, Hammel, G. The core competencies of the corporation. *Harvard Business Review,* May-June, 1990.

3 Quinn, J B, Baruch, J J, Zien, K A. *Innovation explosion*. The Free Press, 1997.

4 Stabell, C B, Fjeldstad, Ø D. Configuring value for competitive advantage : on chains, shops and networks. *Strategic Management Journal,* 19, 1998, 413–437.

5 Evans, P B, Wuster, T S. Strategy and the new economics of investments. *Harvard Business Review*, Sept-Oct, 1997.

6 Løwendahl, B, Revang, Ø. Challenges to existing strategy theory in a post-industrial society. *Strategic Management Journal,* 19, 1998, 755–773.

7 de Bono, E. *The use of lateral thinking*. Pelican, 1971.

8 Pfeffer, J. Six dangerous myths about pay. *Harvard Business Review,* May-June, 1998.

9 Casti, J. *Alternate realities : mathematical models of nature and man.* Wiley, 1989.

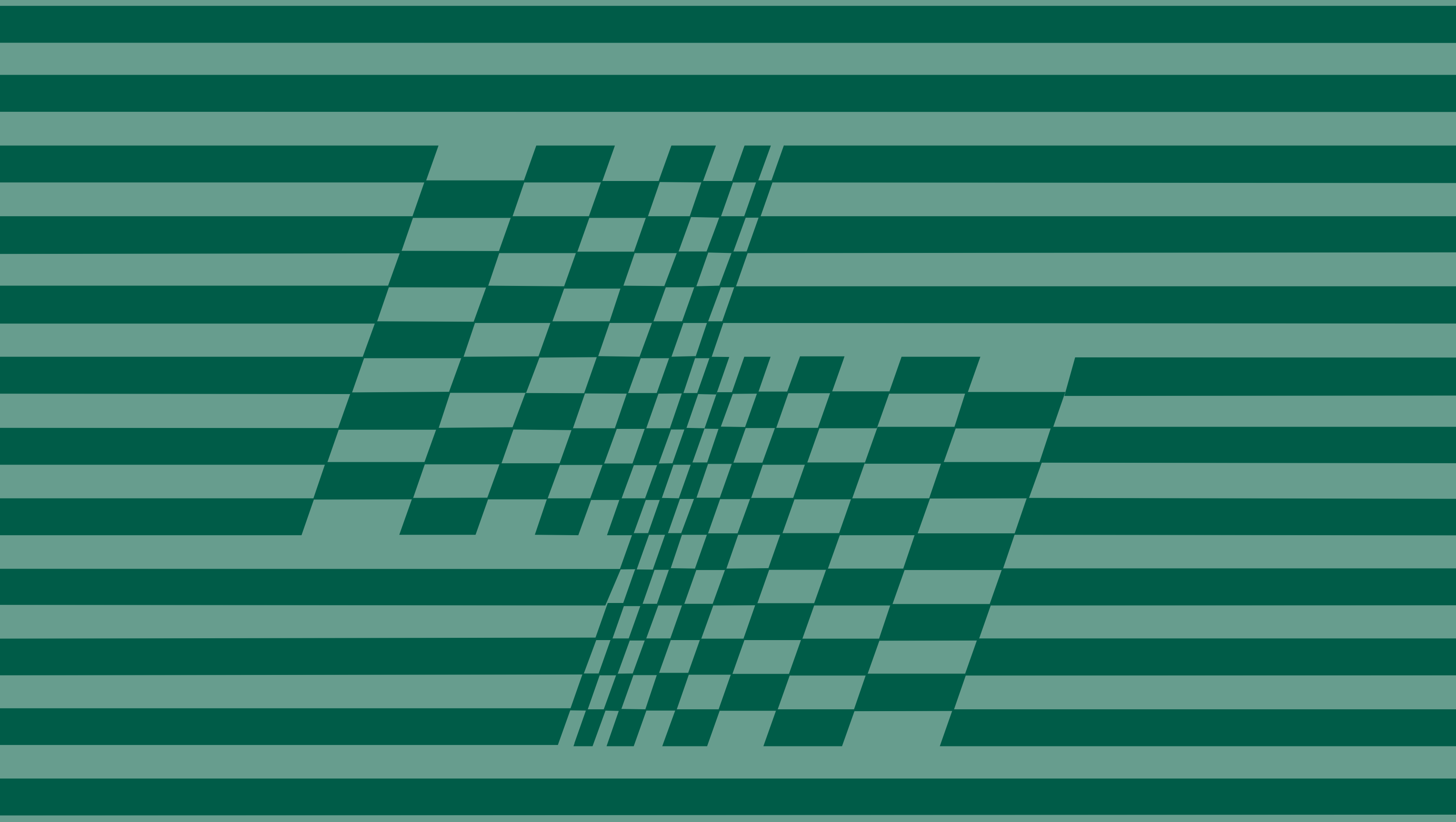
10 Poston, T, Stewart, I. *Catastrophe theory and its applications*. Pitman, 1978.

11 Chandy, R K, Tellis, G J. Organizing for radical product innovation : the overlooked role of willingness to cannibalize. *Journal of Marketing Research,* 35, 1998, 474–487.

Jan A. Audestad (56) is Senior Advisor for the Corporate Management of Telenor. He is also Adjunct Professor of telematics at the Norwegian University of Science and Technology.

e-mail: Jan-Arild.Audestad@s.hk.telenor.no

Special



Voice quality under ATM cell loss

JENS BERGER AND MECHTHILD STOER

This study tries to assess the effect of ATM cell loss and AAL packet loss on voice connections. To this aim, cell losses on a 64 kbit/s CBR stream through one ATM link were measured under high load, and the resulting cell loss trace merged off-line with recorded speech files, simulating AAL1 and AAL5 protocols. We study the impact of different cell stuffing techniques and of packet loss for packets of different lengths. Delay and delay variations have not been measured nor introduced into the speech files. Voice quality was assessed by a formal auditory ('subjective') test and by several instrumental ('objective') methods: PSQM, PSQM+ and DT-SQE. This gives rise to comparisons between cell loss rates, auditory evaluation and instrumental estimation of speech quality.

This work was financed by EURESCOM Project P603 and by Telenor Research Project KN0401.

1 Introduction

The background for this study was the selection and definition of important Quality of Service parameters and the development of measurement methods. This was the scope of EURESCOM project P603 'Quality of Service: Measurement Method Selection', see [3] and [4]. It was found that voice quality still is an important parameter, especially with emerging new technologies for transfer of voice and data. The methods chosen to measure voice quality were:

- A formal auditory test following ITU-T recommendation P.800 [9]
- The instrumental methods
  - 'Perceptual Speech Quality Measure' (PSQM) recommended for codec evaluation in ITU-T P.861 [11]
  - PSQM+ (a modification of PSQM to assess bit error conditions)
  - Deutsche Telekom – Speech Quality Estimation (DT-SQE) [2] designed to assess codecs.

All in all this investigation contained more than 50 different configurations for speech transmission:

- Voice transmission via 'real' telephone networks (ISDN, POTS, GSM)
- Transmission via simulated connections
  - Codecs
  - ATM connections
  - Background noises
- Reference system MNRU (see ITU-T P.810 [10]).

In this paper, we concentrate on the application of these methods to ATM cell loss. This gives rise to interesting comparisons between cell loss, auditory and instrumental speech quality evaluations.

It turns out that up to 2.5 % cell loss inside active speech intervals still produces fair quality, but that the same cell loss rate leads to intolerable quality degradation when cells are packed into larger packets, i.e. when the loss of one cell may lead to the loss of a larger packet. We assume that packet sizes have an impact on voice quality also for voice over IP. It also turns out that PSQM+ and DT-SQE values show a high correlation with the outcome of the auditory test. This gives hope for methods for instrumental voice quality evaluation being applicable for studies and laboratory experiments.

The impact of ATM cell loss on voice quality has also been studied by Meky and Saadawi [12]. They used artificial distributions (uniform, Poisson, etc.) for the cell loss process, and a specially designed instrumental method to evaluate the speech quality. This method is also reported to have good correlation with subjective evaluation.

In our paper we used, instead of artificial cell loss distributions, actually measured cell losses in an ATM switch with high background load and small buffer. Cell losses do not occur evenly spread but in bursts. Since the instrumental methods were designed for evenly spread quality degradations, it was not clear from the beginning that they would also estimate 'correctly' quality degradations occurring in bursts.

A more detailed version of this paper is found in [5].

Section 2 describes how speech files impaired by cell loss were produced. Section 3 describes the evaluation methods consisting of the auditory test and the instrumental methods, and sections 4 and 5 contain results and conclusions.

2 Measurement set-up and test conditions

An ATM cell stream simulating a 64 kbit/s speech connection was fed together with a background load into an ATM switch. The switch is a special designed measurement and experiment switch [13]. The switch was configured so cell loss occurred in a buffer of length 70 toward a 155.52 Mbit/s link. The background load came from synthetic traffic generators (STG) [6]. For a given background load, we registered over a certain time the lost cells of the primary stream. Segments of the cell loss trace were then merged with one of several digital speech files. Thus we produced several impaired speech files. Some more details on the primary 64 kbit/s stream, the background traffic, and the impairment process follow.

2.1 Description of the primary stream

At Telenor Research, we had at hand an automatic traffic generator (Alcatel 8643 ATM Traffic Generator Analyzer, ATGA, see [1]) that is able to generate ATM cells at constant bit rate, with a given virtual channel and virtual path number, with sequence numbering in the payload, and otherwise empty payload. We used these features to emulate a 64 kbit/s user data stream representing the speech connection. AAL1 is intended for transfer of voice connections, see [14]. This protocol reserves 47 bytes of the ATM cell's payload for user data. In the ATGA's cell generation program we set the user's payload to 47 bytes and the rate to 64 kbit/s. This corresponds to a rate of 170.21 cells per second.

Table 1 Cell loss traces

trace name	A	C	E
cell loss on primary stream	$1.125 * 10^{-3}$	$1.097 * 10^{-3}$	$0.879 * 10^{-3}$
measuring time (min)	78	70	74
scenario	HQTV-dominated	CBR-dominated	HSD-dominated

2.2 Description of the background traffic

For generation of the background traffic we used the Synthetic Traffic Generator (STG) developed by Wandel & Golterman, Telenor Research, and SINTEF Tele&Data. A high-level description of its possibilities and its interface can be found in [6]. Source types can be specified inside the STG as Markov models with state sojourn times, state transition probabilities, and states corresponding to the sending of a predefined cell pattern. We used three types of sources: constant bit rate traffic (CBR), high speed data traffic (HSD), and high-quality television (HQTV), whose Markov source models were defined in [7]. From these source types three traffic scenarios are set together:

- An HQTV-dominated traffic scenario where the load from the sources of the three types (HQTV, HSD, CBR) contributes approximately in the ratio 0.6 : 0.2 : 0.2;
- An HSD-dominated traffic scenario where the load from the sources of the three types (HQTV, HSD, CBR) contributes approximately in the ratio 0.2 : 0.6 : 0.2;
- A CBR-dominated traffic scenario where the load from the sources of the three types (HQTV, HSD, CBR) contributes approximately in the ratio 0.2 : 0.2 : 0.6.

The total mean load of each of these traffic mixes was set such that the total stationary cell loss ratio becomes approximately 10^{-3}. This high value (instead of the more usual 10^{-9}) was chosen, because we want to look at the effect of the otherwise rather seldom cell loss bursts. Another reason for this choice of cell loss ratio was that the speech file segments to be impaired were maximally 8 seconds long. This corresponds to about 1362 cells.

Table 1 shows the cell loss rates (cells lost relative to total number of cells sent) on the primary stream for measuring times of about one hour, for each scenario. We did not measure the total cell loss rate of the primary stream together with the background streams.

2.3 Merging the cell loss pattern with the speech files

The available speech files were coded in linear PCM with sampling rate 8 kHz and 2 bytes per sample. They consisted of segments 8 second long. Each cell loss trace file was split into non-overlapping windows of 8 seconds, which corresponds to 1362 sent cells of the primary stream.

We simulated packing of a speech file into 1362 cells (with sample 1 into cell 1), where 47 samples would fit into 1 cell, as the AAL1 protocol would do. There is one difficulty here: The cell payload is 47 bytes, but 47 samples of the original speech files consist of *twice* 47 bytes. A more proper simulation procedure would have been to (sample-wise) code the speech files using A- or μ-law, thus halving the file size, then to simulate the packing of coded samples into cells, and to decode the files again. We skipped the coding/decoding procedure, since quality reduction by A- or μ-law coding is negligible.

There are several ways to replace samples in lost cells:

1. Silence insertion: The lost 47 samples are replaced by 47 samples of value 0, which denotes silence in linear PCM.
2. Repeating: The lost 47 samples are replaced by the 47 samples in the last received cell. If the first cell was lost, its samples are replaced by silence.
3. Packet replacement: Here we simulate the sending of the file in packets of *k* cells. Whenever one or more cells are lost in a packet, the whole packet is lost. This simulates the AAL5 protocol for sending of packets. Samples in lost packets are replaced by silence-value 0. We used the arbitrary values of 3 and 6 for *k*.

To be exact, the user payload under the AAL5 protocol is on the average between 47 and 48 bytes, but we ignored the exact values.

The three replacement methods are abbreviated ‘s’ for silence replacement, ‘r’ for repeating, and ‘s3’ or ‘s6’ for replacement of packets of 3, resp. 6, cells by silence.

2.4 Conditions tested

For the auditory test, five German language files of 32 seconds length each were prepared, which were to be impaired under five different conditions. The 32-second files contained four 8 second segments spoken by two male and two female speakers. The 8 second segments each contained two whole sentences by one speaker.

An example condition in the ATM case is “each of the four 8 second segments of a 32 second file is impaired by loss of *x* packets of size *k* of the xxx-dominated scenario; and the loss occurs in a period of active speech.” By varying the number of packets *x* and their size *k*, we received five conditions.

We produced the impairments in the following way. From one of the cell-loss traces A to E we chose four 8 second windows, in which about *x* packets of size *k* were lost by the packet replacement method described in Section 2.3. The loss of *x* packets usually occurred in bursts of maximal length 70 ms. A difficulty was to place the burst inside a period of active speech of an 8 second segment. We tried, inside one 32 second file, to make one burst meet the end of a sentence, another the beginning of a sentence, and the rest the middle of a sentence.

Table 2 shows the conditions. The trace name refers to Table 1 (cell loss traces), and the replacement method refers to section 2.3.

These five conditions were included in a set of altogether 53 different conditions. Therefore the results are directly comparable with the considered references and the other conditions in the test. The speech quality was evaluated in an auditory test performed at Deutsche Telekom Berkom. The quality of the impaired speech file was also evaluated by PSQM [11], PSQM+ (a modification of PSQM) and DT-SQE [2] as a function of time. In the next section we describe the auditory test and the instrumental methods.

Table 2 Impaired files for auditory test

impaired file	trace	packets lost in 8 sec segment (x)	packet size (k cells)	replacement method
g30_atm	A	11–14	1	s
g31_atm	A	12–14	1	r
g32_atm	C	7–8	3	s3
g33_atm	C	7	6	s6
g34_atm	E	11–12	6	s6

g30_atm and g31_atm are using the same cell loss patterns.
So are g32_atm and g33_atm.

3 Evaluation methods

3.1 Description of the auditory test

The best model for evaluation of speech transmission quality would be a set of normal free conversations between two test participants. Such kinds of tests make it possible to consider and to evaluate all characteristics that influence the speech quality (or better, the capability for conversation) in an end-to-end connection. But conversational tests with two partners in a test laboratory are very time consuming and require a high number of test participants. Therefore, for estimation of the speech quality of transmission systems (i.e. codecs, filter systems) the usage of Listening-Only-Tests (LOTs) is very common. These tests are well suited for a very efficient estimation of the one-way transmission quality. Some features of quality, which are only noticed in conversations, cannot influence the test results (e.g. signal delay, talker echoes, semi-duplex characteristics). The efficiency of LOTs makes it possible to evaluate a high number of test conditions in one listening experiment using the same test participants and well-prepared speech material for statistical validity. Another advantage of LOTs is the off-line test procedure; in the test recorded speech samples are used. There is no need to have the current telephone connection available for the auditory test.

Design of the auditory test

Test participants

- The group of test participants consists of typical non-technical subjects selected from a pool of listeners used for auditory tests.
- The age and the educational background of the subjects provide a good representation of the typical telephone user population.
- The percentage of female and male listeners is 50 % each.
- All test participants were previously audiometrically checked for normal hearing threshold by pure tone audiometry.

Speech material

Each 32 second test file was split into four 8-second speech samples. The used speech samples consist of two short German sentences coupled by a silence interval spoken by one person. The whole 32 second file is spoken by two males and two females. All samples are selected from a phonologically balanced speech database.

Scoring

In accordance with ITU-T P.800 [9] a five-step Absolute Category Rating (ACR) scale is offered for scoring:

5	excellent
4	good
3	fair
2	poor
1	bad

All test participants (24 in our case) score all conditions in the test. The 53 tested ‘conditions’ correspond to 53 different 32 second speech files. Five of these were impaired by ATM cell loss, as described in the previous section, the other conditions involved ISDN, GSM, POTS connections, and different codec simulations. A full description can be found in [4].

After each offered speech sample (length: 8 sec) each test listener has to score the perceived overall quality. The speech transmission quality will be described by a Mean Opinion Score (MOS) for each condition. The MOS value is here defined as the mean value of all individual results of the test listener and the four speakers (in total: 96 single results for every condition).

3.1 Methods for instrumental assessment

Instrumental approaches to speech quality prediction try to achieve a high correlation between the predicted quality values and the scores gained by auditory tests. Most instrumental methods for speech quality estimation compare the (undistorted) source speech signal and the impaired output signal of the transmission system.

A great number of instrumental approaches work in several steps. The first step usually eliminates signal differences that are irrelevant in the modelled auditory test (for example total delay and level differences). The next stage transforms both signals to an ‘internal representation’ using psycho-acoustic models for the human sound perception. The spread (including multidimensional aspects) between both pre-processed signals is computed and will be used for estimating a quality value. In this way modern instrumental approaches based on psycho-acoustic and cog-

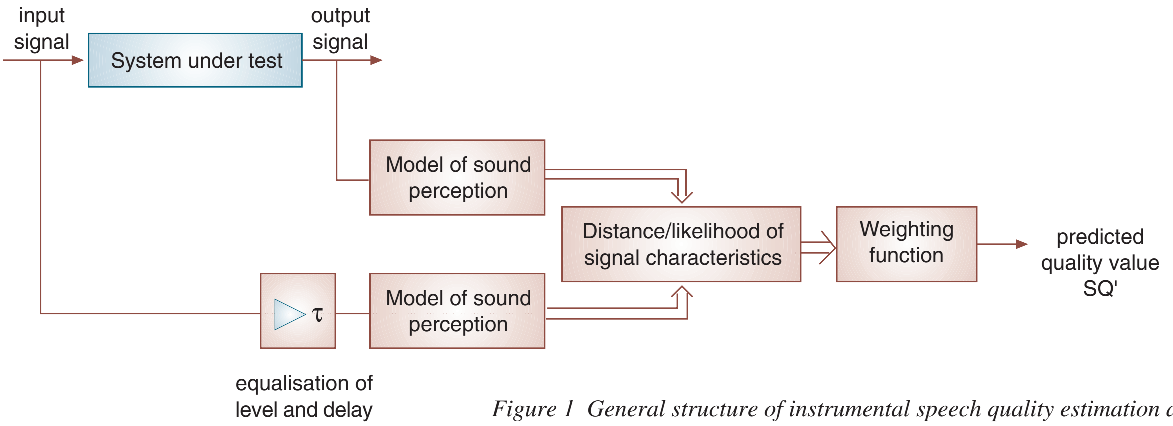

input signal
System under test
output signal
Model of sound perception
Distance/likelihood of signal characteristics
Weighting function
predicted quality value SQ'
τ
Model of sound perception
equalisation of level and delay

Figure 1 General structure of instrumental speech quality estimation approaches

nitive knowledge are models of the auditory test with human test listeners (Figure 1).

Several approaches or methods are known which differ in e.g.:

- The kind of signal pre-processing;
- The used psycho-acoustical models;
- The determination of a value describing the speech quality; and
- The database for definition of the fitting function to transform the technical speech quality value to a predicted MOS value.

DT-SQE has been designed in the Deutsche Telekom Berkom research group 'Quality and Acceptance of Voice Services' (former: 'Speech Quality Measurement, Psycho-acoustics' Research Centre of Deutsche Telekom AG). A former approach of DT-SQE was published in connection with an investigation of DCME systems in [2]. The current version of DT-SQE used in EURESCOM project P603 was tuned using German speech samples taken from ITU-T and ETSI speech quality tests. This approach will be invited in ITU named as T-OSQA.

PSQM (Perceptual Speech Quality Measure) is a method for measuring the quality of narrowband (300 – 3400 Hz) speech codecs. It is recommended by ITU-T P.861 [11]. *PSQM 0.0* is a version of PSQM in which distortions in silence intervals are weighted with 0.0 (i.e. not considered at all). *PSQM 0.2* weights silence intervals with 0.2. Silence weight 0.2 is recommended by ITU-T P.861 [11]. *PSQM+* is an experimental version of PSQM designed for better handling of bit-errors in GSM.

PSQM and DT-SQE have been developed and optimised for evaluating and rating speech codecs. Only limited experience has been gained by experts using these methods to assess end-to-end quality of international (inter)connections or to assess ATM or IP connections.

Features of the used instrumental approaches

- DT-SQE (Deutsche Telekom – Speech Quality Estimation) and PSQM (Perceptual Speech Quality Measure) follow this common structure and components as shown in Figure 1.
- The first step in both measures is the total delay and gain adjustment (these values are given or computed before).
- DT-SQE and PSQM consider the telephone band limitation and the receiving function of an 'ordinary' handset.
- Both measures are based on a short-time spectral analysis (t_W = 16 … 20 msec) and compute the levels in the critical bands for each signal and each segment (frequency warping).
  - PSQM divides the spectrum into 0.25 Bark bands; DT-SQE only into 1 Bark band.
  - Additionally, DT-SQE reduces the influence of total frequency shape in this step.
  - PSQM adds a Hoth noise to both signal representations.
- Intensity warping of the critical band levels:
  - DT-SQE is based strongly on the *Zwicker* model of calculation of specific loudness. Spectral masking and the compression are taken from the Zwicker model [15].
  - PSQM considers masking effects only in a simplified way and uses a much higher compression than *Zwicker.*
- PSQM calculates a 'loudness scaling factor' to adjust the loudness in the current segment.
- DT-SQE reduces differences between both signal representations, which have nil or a restricted influence in the auditory test. One part of this step is also a slight asymmetry processing (see PSQM).
- Both modified short time loudness patterns are used for calculation of a quality value in the current segment.
- The computed similarity is the main result of measuring the speech quality in DT-SQE. The mean value of all short time results yields the internal speech quality value.

- In PSQM the result of a computed ‘cognitive subtraction’ after an ‘asymmetry processing’ describes the current quality value.

It is often desirable to express voice quality in MOS values, but it is difficult to determine a unique function which transforms the DT-SQE or the PSQM values to the estimated MOS value. This is because absolute MOS values depend on the context of the auditory tests. Also cultural differences or differences in language can have significant effect on the auditory MOS. Therefore, we decided in this study to use the internal speech quality values directly. For DT-SQE a mapping function was used, which was optimised for the ITU-T 8 kbit Test, 1994. But in the context of this test the results should only be used as ‘quality describing values’, not as estimated MOS.

4 Results of the analysis

In this section we compare the subjective MOS values with the instrumental values for the files impaired by ATM cell loss. Table 3 shows the results for the 8-second segments and Table 4 the results for the 32-second files composed of the 8-second segments. One should note that the subjective MOS in Table 3

Table 3 Results from auditory test for 8-second segments

segment	subj. MOS	PSQM 0.0	PSQM 0.2	PSQM+ 0.0	DT-SQE-MOS
g30_atm.1	3.292	0.079	0.075	0.083	3.959
g30_atm.2	3.125	0.528	0.301	0.593	3.924
g30_atm.3	2.792	0.869	0.570	0.993	3.726
g30_atm.4	3.042	0.410	0.266	0.477	3.970
g31_atm.1	3.667	0.218	0.158	0.215	3.829
g31_atm.2	4.292	0.113	0.090	0.114	3.964
g31_atm.3	4.667	0.213	0.132	0.206	3.975
g31_atm.4	3.458	0.695	0.470	0.697	3.779
g32_atm.1	2.750	0.211	0.169	0.872	3.487
g32_atm.2	2.083	0.618	0.389	1.201	3.777
g32_atm.3	2.708	0.451	0.340	1.188	3.303
g32_atm.4	2.750	0.413	0.322	0.779	3.918
g33_atm.1	2.250	0.371	0.242	2.088	3.080
g33_atm.2	2.208	0.381	0.282	1.460	3.063
g33_atm.3	2.708	0.646	0.485	2.992	3.152
g33_atm.4	2.625	0.519	0.397	2.685	3.239
g34_atm.1	2.083	0.535	0.367	3.923	2.809
g34_atm.2	2.417	0.645	0.396	4.656	2.234
g34_atm.3	1.792	0.951	0.557	7.399	2.392
g34_atm.4	1.833	0.967	0.693	4.666	2.699

Table 4 Results from auditory test for 32-second files

segment	subj. MOS	PSQM 0.0	PSQM 0.2	PSQM+ 0.0	DT-SQE-MOS
g30_atm	3.062	0.483	0.312	0.55	3.849
g31_atm	4.021	0.326	0.219	0.32	3.873
g32_atm	2.573	0.411	0.306	0.99	3.637
g33_atm	2.448	0.493	0.365	2.39	3.055
g34_atm	2.031	0.773	0.509	5.04	2.634

is averaged over 24 observations only. The 95 % confidence intervals are of size 0.24 to 0.52. So these values have to be treated carefully. In contrast, the subjective MOS values in Table 3 have been averaged over 96 observations. The 95 % confidence intervals for the subjective MOS values in Table 4 are smaller, of size 0.18 to 0.23.

4.1 Comparison of instrumental values with subjective MOS

In Figure 2 to Figure 5 the instrumental values are plotted against the subjective MOS for each of the 8-second segments.

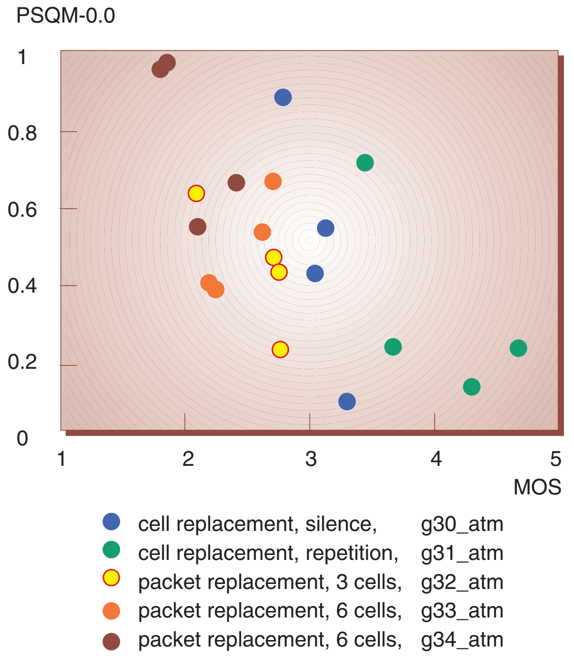


Figure 2 PSQM 0.0 versus subjective MOS. 8-second segments

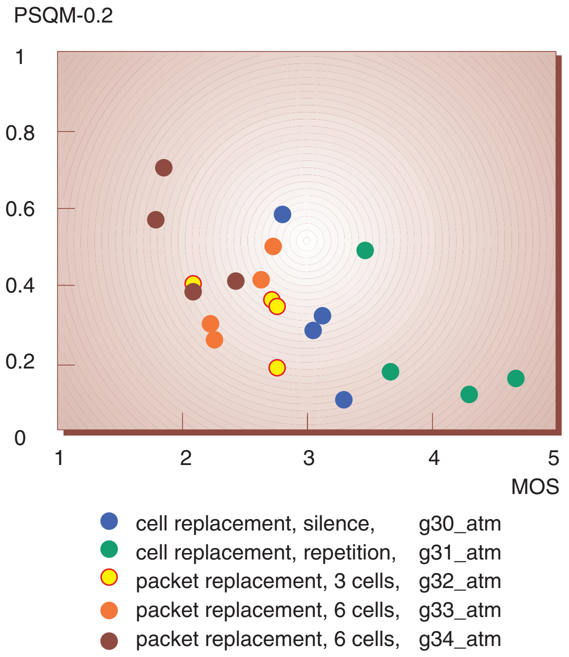


Figure 3 PSQM 0.2 versus subjective MOS. 8-second segments

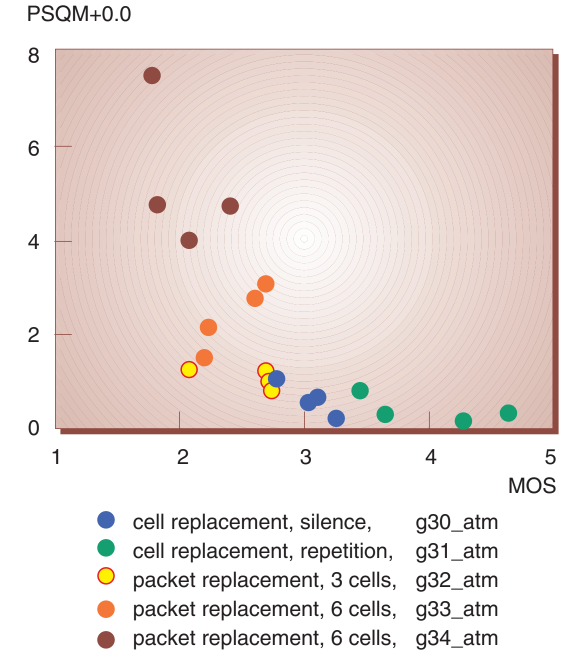


Figure 4 PSQM+0.0 versus subjective MOS. 8-second segments

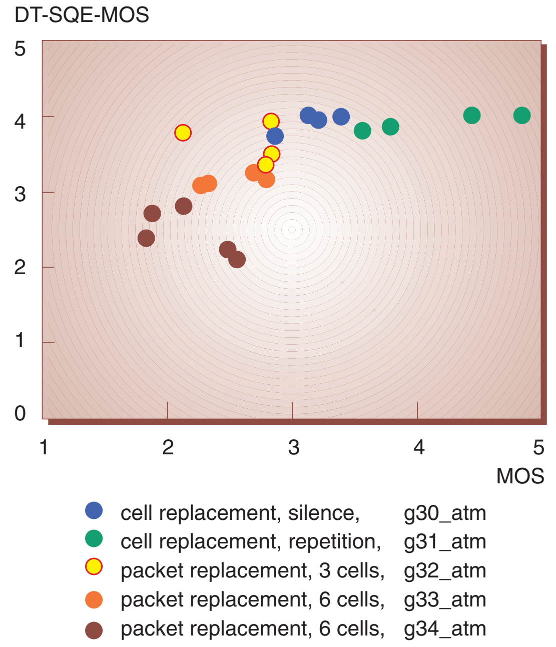


Figure 5 DT-SQE versus subjective MOS. 8-second segments

Table 5 shows the correlation between the subjective MOS and a linear or quadratic fitting function in the respective instrumental values.

Even taking into account the unreliability of the subjective MOS data for the 8-second segments, it seems that PSQM+ and DT-SQE have a better correlation with the subjective MOS than PSQM has, no matter which silent weight is used.

It does not really make sense to compute correlations of subjective MOS with the instrumental values for the 32-second files, because there were only five files. Instead, we plot the results in Figure 6 to Figure 9.

There are a few things to be noted here. First, the instrumental methods make almost no difference between the two cell replacement methods: silence or repetition of last cell content (see g30_atm and g31_atm). But the subjective impression is that the repetition method gives better results than replacement by silence. We do not know whether our data allow strong conclusions to be drawn about the quality of cell replacement methods.

As was to be expected, dropping long packets leads to higher quality reductions than dropping single cells. PSQM+ and DT-SQE properly register this fact, but PSQM has difficulties.

The transformation of PSQM values to MOS values given in [4] is useless. As can be seen from Figure 10, the estimated MOS values range only from 3.95 to 4.05 for all files. MOS value 4 corresponds to 'good' quality. This does not mean the PSQM method is useless, but that the transformation of PSQM values into MOS needs to be redesigned in the case of ATM cell loss. MOS estimated from PSQM+ is not quite linearly correlated with subjective MOS either, but can be used as an approximation.

Table 5 Correlation of MOS to instrumental values for 8-second segments

instrumental value	linear	quadratic
PSQM 0.0	0.63	0.65
PSQM 0.2	0.64	0.68
PSQM 0.4	0.67	0.72
PSQM+ 0.0	0.70	0.77
PSQM+ 0.2	0.71	0.77
PSQM+ 0.4	0.72	0.78
DT-SQE	0.71	0.76

For the right classification of the results Figure 10 shows the auditory MOS values versus estimated MOS' values gained by the instrumental approaches for all 53 conditions in the test. For this reason the instrumental speech quality values are mapped into the MOS scale using a monotonous range of an optimal quadratic function. The five ATM conditions are drawn as black filled squares.

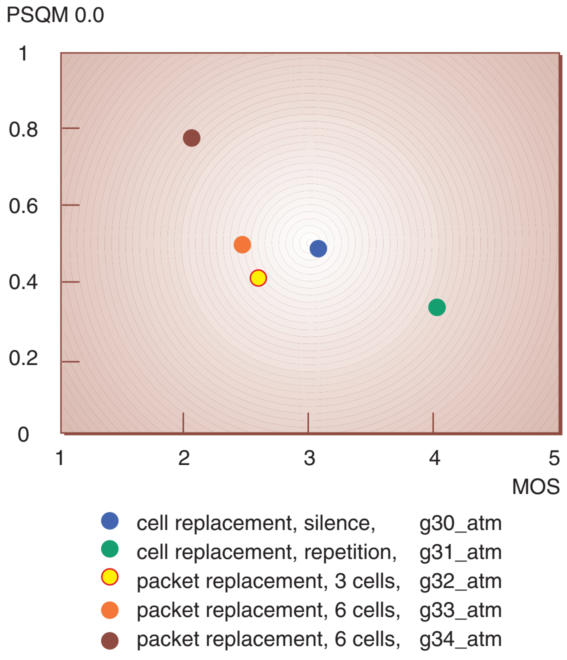


Figure 6 PSQM 0.0 versus subjective MOS. 32-second files

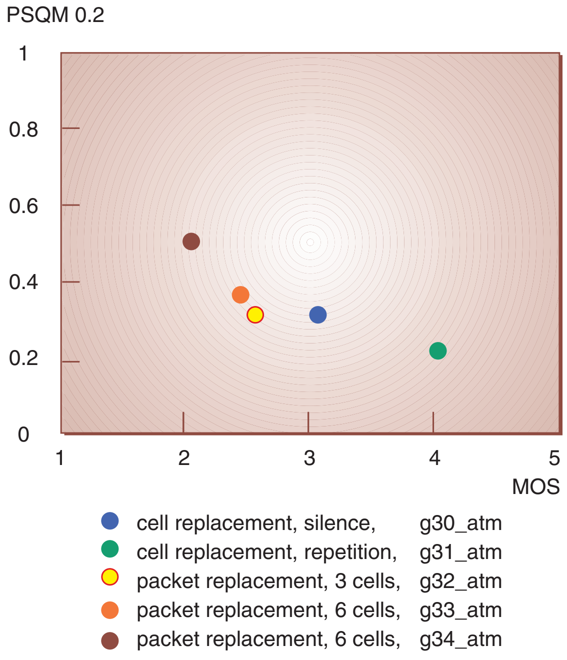


Figure 7 PSQM 0.2 versus subjective MOS. 32-second files

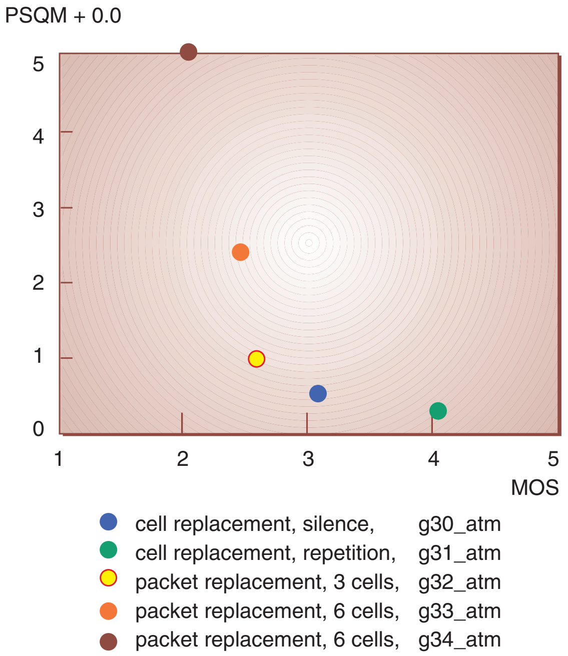


Figure 8 PSQM+0.0 versus subjective MOS. 32-second files

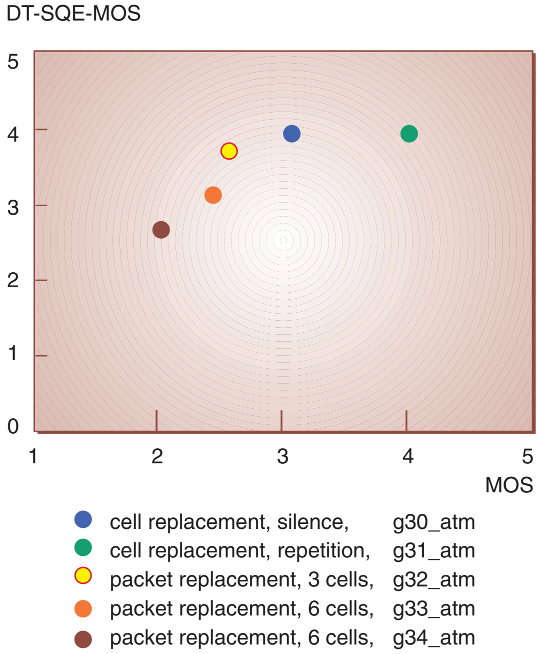


Figure 9 DT-SQE versus subjective MOS. 32-second files

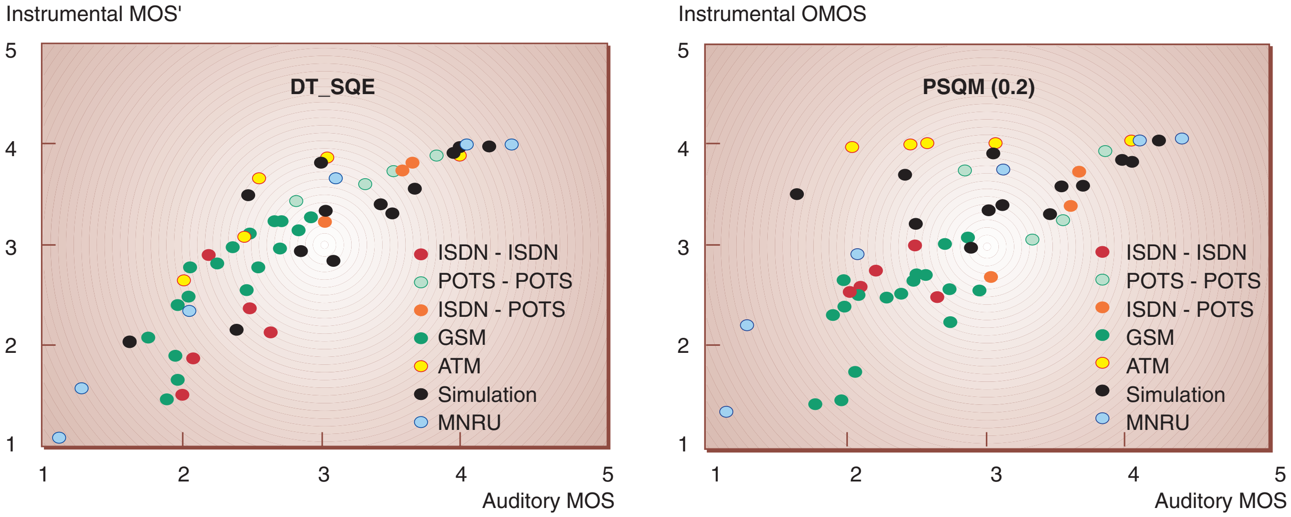


Figure 10 Results gained by DT-SQE and PSQM(0.2) versus subjective MOS

4.2 Comparison of subjective MOS with cell loss and packet loss rates

Does the cell loss rate have any bearing on the subjective impression of quality reduction, or does packet loss have more to say? In Figure 11 the *sample loss rate* in *active speech intervals* is plotted against the subjective MOS. This corresponds to the packet loss rate (packets lost in active speech intervals divided by packets sent in active speech intervals, with an adjustment for packets on the border of speech intervals). The definitions of active speech intervals and of silent intervals are taken from PSQM, see [11]. The exact computation of sample loss rate is described in [5].

In Figure 12 the *cell loss rate* in active speech intervals is plotted against subjective MOS. The cell loss rate refers only

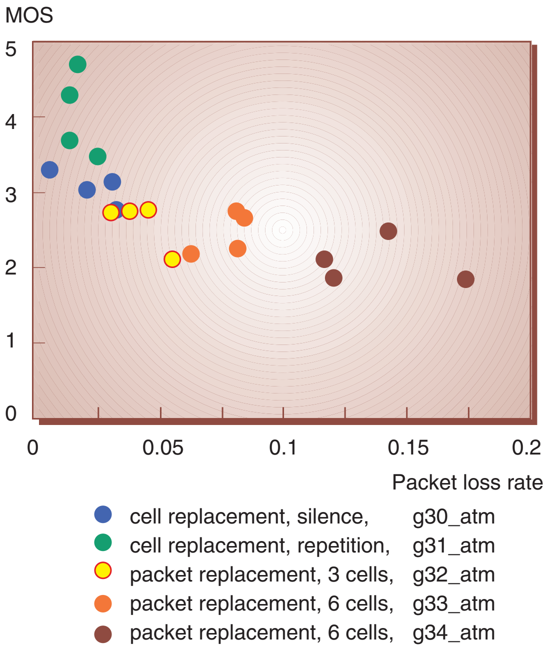


Figure 11 Subjective MOS versus packet loss rate in active speech

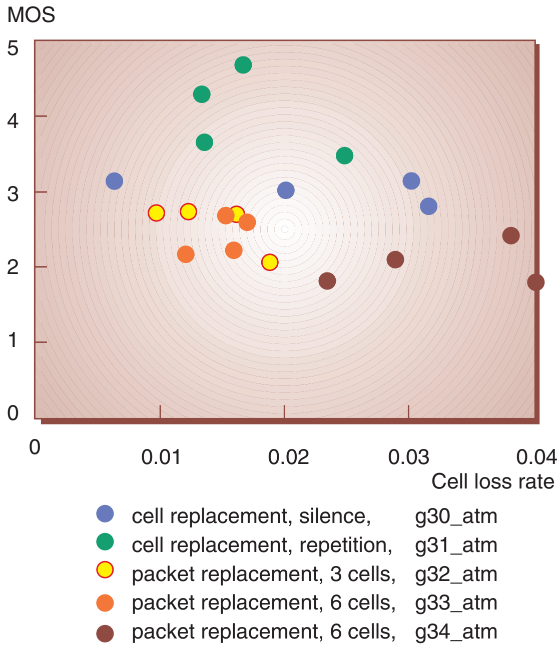


Figure 12 Subjective MOS versus cell loss rate in active speech

to cells lost on the ATM level (i.e. the cell loss trace), not those lost inside packets on the ATM adaptation layer level.

As was to be expected, the cell loss rate (inside active speech) has almost nothing to say on the speech quality. One has to take into account how the protocols above the ATM layer pack cells into packets, and what happens when cells get lost.

One may also conclude that if the cell loss rate is maximally 0.025, the quality of cell-impaired speech is ‘fair’, depending on the cell replacement method.

5 Conclusions

Conclusions for voice over ATM

Judging from eight speech samples heard by 24 listeners, the quality of speech impaired by cell loss (not packet loss) is still fair, when the cell loss rate is about 0.025. Replacing lost cells by the contents of the last received cell gives a somewhat better quality than replacing them by silence.

The sample loss rate (computed from the number of samples in a lost packet) has more to say on speech quality than the cell loss rate, which ignores packet sizes.

The longer the packet size, the more serious the effect of cell loss in the underlying ATM layer. This may as well hold true for IP-over-ATM connections.

Conclusions on the instrumental methods

When comparing the different instrumental measures with the subjective MOS, PSQM+ and DT-SQE show a better correlation than PSQM. Although these methods were not designed to evaluate quality impairments caused by packet losses, PSQM+ and DT-SQE perform quite well for these conditions.

The instrumental methods possibly need to be improved to estimate differences in speech quality, when different cell replacement techniques are used. Subjective listening experience shows that replacing lost cells by the content of the last received cell gives better quality than replacing lost cells by silence samples (the difference in quality is as large as from ‘fair’ to ‘good’), whereas all instrumental methods judge the quality of the two replacement methods to be roughly equal.

Limitations to this study

This work has a few shortcomings, which could be the topic of further work.

- Cell delay and delay variation was not a part of the study.
- Only impairments to a single 64 kbit/s channel were investigated. Encoding to lower bitrates, or multiplexing of several channels into one ATM cell stream were not part of the study.
- Reconstruction of lost packets or cells may have a beneficial effect on the voice quality.
- Newer ATM switches give priority to CBR traffic such as traffic from voice connections. Thus it is expected that cell loss will be a very rare event in a well-dimensioned net. However, when voice traffic creates a substantial part of the load, cell loss will play a role.

Acknowledgements

We express our thanks to all members of EURESCOM project P603, especially to Bjarne E. Helvik for helpful discussions, and to Michel Emons for providing us with PSQM evaluations of impaired speech files. Moreover, we thank Egil Aarstad for configuring the ATM equipment.

6 References

1 *Alcatel 8643 ATGA User Manual.* Zürich, Alcatel STR AG.

2 Berger, J. Ein Ansatz zur instrumentellen Sprachqualitätsabschätzung im Festnetz der Deutschen Telekom. (An approach to instrumental speech quality estimation in the Deutsche Telekom's fixed network). *Workshop on quality assessment in speech, audio and image communication, ITG, EURASIP, Darmstadt,* March 1996. (In German.)

3 EURESCOM. *Quality of Service : Measurement Method Selection.* EURESCOM P603, Deliverable 1, QoS overview, services and measurement methods : selected services and methods, 1997.

4 EURESCOM. *Measurement method.* EURESCOM P603, Deliverable 2, 1997.

5 EURESCOM. *ATM Measurement Report.* EURESCOM P603, Deliverable 2, Annex 1, 1997.

6 Helvik, B E. Synthetic load generation for ATM traffic measurements. *Telektronikk,* 91, (2/3), 174–194, 1995

7 Helvik, B E, Stol, N. *ATM traffic experiments : a laboratory study of service interaction, loss fairness and loss characteristic.* Trondheim, SINTEF-DELAB, 1995. (Report STF40 A95024.)

8 ITU-T. *B-ISDN ATM adaptation layer specification : type 1 AAL.* ITU-T Recommendation I.363.1, 1996.

9 ITU-T. *Methods for subjective determination of transmission quality.* ITU-T Recommendation P.800, 1996.

10 ITU-T. *Modulated noise reference unit (MNRU).* ITU-T Recommendation P.810, 1996.

11 ITU-T Recommendation P.861. Objective quality measurement of telephone-band (300–3400 Hz) speech codecs, 1996.

12 Meky, M M, Saadawi, T N. Degradation effect of cell loss on speech quality over ATM networks. In: Mason, L and Casaca, A (eds.). *Broadband Communications.* London, Chapman Hall, 1996, 259–270.

13 Myrstad, T. Test-svitsj for bredbånds ISDN (Test switch for broadband ISDN). *Telektronikk,* 88, (3), 63–65, 1992. (In Norwegian.)

14 Wright, D J. Voice over ATM : An evaluation of implementation alternatives. *IEEE Communications Magazine,* 34, (5), 72–80, 1996.

15 Zwicker, E, Fastl, H. *Psychoacoustics, facts and models.* Berlin, Heidelberg, Springer-Verlag, 1990.

Abbreviations

AAL	ATM Adaptation Layer
ATGA	ATM Traffic Generator Analyzer, Alcatel 8643
ATM	Asynchronous Transfer Mode
CBR	Constant Bit Rate
DT-SQE	Deutsche Telekom – Speech Quality Estimation
HQTV	High Quality Television
HSD	High Speed Data
MOS	Mean Opinion Score
PSQM	Perceptual Speech Quality Measure
STG	Synthetic Traffic Generator

Jens Berger is Research Scientist in Quality and Acceptability of Tele-Services, a branch of Deutsche Telekom Berkom GmbH in Berlin. He started off in the area of acoustics and audio sound systems but is currently involved in national and international projects in speech quality evaluation for telecommunications. His Ph.D. thesis from 1998 focuses on instrumental ('objective') approaches for speech quality estimation.

e-mail: j.berger@berkom.de

Mechthild Stoer (35) is Research Scientist at Telenor Research and Development. She holds a Ph.D. degree from the University of Augsburg 1991 in the area of combinatorial optimization. Her interests are in network design and network quality issues.

e-mail: mechthild.stoer@fou.telenor.no

Traffic from Mobility in Mobile Broadband Systems

FERNANDO J. VELEZ AND LUIS M. CORREIA

Models allowing the study of the influence of coverage distance and velocity on the supported traffic and on the new calls traffic linear density are examined, and results are obtained for typical scenarios in a Mobile Broadband System (MBS) with a linear coverage geometry. For systems without guard channels for handover, for a fixed bounding value for the blocking probability, the new calls traffic linear density was analyzed, increasing with the decrease of the maximum coverage distance, *R*, being upper limited by a value which depends on the characteristics of the mobility scenario. However, call-dropping probability requirements also need to be fulfilled, leading to a new calls traffic density that only increases with the decrease of *R* down to an optimum value of *R*, and being lower for scenarios with higher mobility. These optimum values of *R* are higher for scenarios with higher and higher mobility, leading to limitations in system capacity, mainly for high mobility scenarios. In order to resolve these limitations, the use of guard channels for handover is studied, particularly for high mobility scenarios. For these scenarios one concludes that there is a degradation in system capacity because, for the typical coverage distances foreseen for MBS, the new calls traffic linear density is one order of magnitude below the values obtained for the pedestrian scenario (where it is approximately 15 Erlang/km), decreasing from 2.47 Erlang/km, in the urban scenario, down to 0.84 Erlang/km, in the highway scenario, when two guard channels are used.

1 Introduction

Mobile Broadband Systems will allow extending high data rates provided by the fixed broadband-ISDN to the cellular communications market, supporting high speed communications in high mobility outdoor scenarios [1], leading to the use of millimetre wavebands. For such bands, shadowing from buildings is important, the propagation being mainly in line-of-sight. As a consequence, for urban scenarios the system will be based on a microcellular structure with cells confined to streets, having dimensions in the order of a few hundred metres [2]; also for main road scenarios, the use of microcells is foreseen.

The high mobility associated with the future Mobile Broadband System (MBS) yields a teletraffic analysis where both the new calls and the handover traffic should be considered simultaneously. For systems where guard channels for handover are used, this analysis is made assuming that handover traffic is Poisson distributed [3], and that there is independence among the number of calls being served at each cell [4].

One of the goals of system planning is the maximization of the new calls traffic in terms of the cells dimension, i.e. one is interested in obtaining the cell coverage range that leads to a maximum new calls traffic density supported by the system. For the case of linear geometry, where mobiles travel randomly through cells with maximum coverage length *R*, total length $L = 2R$, and base stations located in the centre, located end-to-end (Figure 1), a new calls traffic per unit length ξ_n (or new calls traffic linear density) is considered. To have an insight into the trade-offs involved in the optimization procedure for MBS, the behaviour of ξ_n in terms of the maximum coverage distance of a cell *R*, needs to be studied for typical values of *R*, and for various mobility scenarios: static, pedestrian, urban, main roads and highways.

At first glance, one could consider that the new calls traffic linear density should be proportional to $(1/R)$; however, this is only valid for the static scenario, where there are no handovers. In scenarios with mobility, the number of calls generated by handover increases as *L* decreases and the velocity increases, implying that ξ_n does not increase linearly with $(1/R)$. For a given supported traffic, the increasing behaviour of ξ_n, associated with the decrease of the coverage distance, corresponds to an increase of the cross-over rate η (number of handovers per unit length) [4]. These facts originate higher probabilities for handover failure P_{hf} and call dropping probabilities P_d with the decrease of *R*.

The simple situation of homogeneous traffic (constant value of new calls traffic in the whole service area) and linear coverage geometry (where mobiles handover between the first and the last cells, a typical geometry for roundabouts [5]) will be considered here as a first step to a more complicated (and closer to reality) analysis. The objective of the design is to obtain values for *R* that verify the requirements of system quality. These requirements consist of values lower than 1 – 2 % for the blocking probability P_b [6] and lower than 0.1 – 0.5 % for the call dropping probability [7].

When no guard channels for handover are used the blocking and the handover failure probabilities are equal, which strongly limits system capacity for high handover failure probabilities, i.e. low coverage distances. If guard channels for handover are used, equations for blocking and handover failure probabilities will be decoupled, and trade-offs can be used to improve system performance in the case of small cells.

Although MBS will be a multi-service system, providing multimedia mobile communications handling both bursty and constant bit-rate traffic, in this work only a single service will be taken into account in order to simplify this first analysis of the problem. A service with mean duration of 3 min is considered as an example. Anyway, this analysis will be very useful in the computation of multi-service aggregated traffic, because the first step of this more general problem [8] consists of computing, for each type of traffic, the state probability as if it was the only kind of traffic in the system.

In Section 2, the influence of traffic from mobility on the microcellular coverage distance is examined. Time parameters are described and the handover probability is introduced. The characteristics of some main mobility scenarios are introduced and formulas for the cross-over rate are then obtained, from which values for the average cross-over velocity are calculated. In systems without guard channels for handover, for given values of the number of channels in each cell and of P_b, one can calculate the corresponding supported traffic. Equations for the new calls traffic and the traffic coming from handover are also obtained.

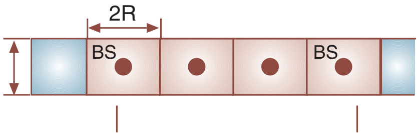


Figure 1 Linear coverage geometry

Models for the computation of new calls traffic linear density are presented, its dependence on *R* is highlighted and the resulting limitations in system capacity imposed because of the high mobility of terminals are discussed.

In Section 3, the use of guard channels for handover is proposed as a prospective solution to overtake the limitations discussed in Section 3 for the new calls traffic linear density. First, the assumptions made in the traffic analysis are described. Finally, the trade-offs involved in the design are described and results are obtained for the supported traffic, the new calls traffic linear density and coverage distances that maximize it.

Finally, in Section 4, some conclusions are drawn on the influence of mobility on the new calls traffic linear density and its consequences on the optimization of MBS capacity.

2 Influence on the Optimum Microcellular Coverage Distance

2.1 Traffic Requirements

In a linear coverage geometry, cells are placed end-to-end and mobiles can handover from a cell only to one of the two adjacent ones, Figure 2; a call comprises successive sessions τ_1, τ_2, τ_3, ... in cells traversed by a mobile terminal, and its duration τ follows an exponential distribution whose mean is $\bar{\tau} = 1 / \mu$ [4], where μ is the service rate. The channel occupancy time τ_c is the time spent by a user in communication prior to handover (or subsequent to handover) or call completion, which can also be modelled by an exponential distribution with reasonable accuracy [9].

The cell dwell time τ_h is the residing time of a mobile within a cell. Further assuming that the dwell time is exponentially distributed with mean $\bar{\tau}_h = 1 / \eta$, then the channel occupancy time is $\tau_c = \min\{\tau, \tau_h\}$, i.e. it is either the time spent in a cell before crossing the cell boundary if the call continues, or the time until the channel is relinquished [4]. As the minimum of two exponentially distributed random variables is also exponentially distributed with parameter $\mu_c = \mu + \eta$, the mean channel occupancy time is given by

$$\bar{\tau}_c = \frac{1}{\mu_c} = \frac{1}{\mu + \eta} \tag{2.1}$$

the probability of handover being given by

$$P_h = \mathrm{Prob}\,\{\tau > \tau_h\} = \frac{\eta}{\mu + \eta} = \frac{\bar{\tau}_c}{\bar{\tau}_h} \tag{2.2}$$

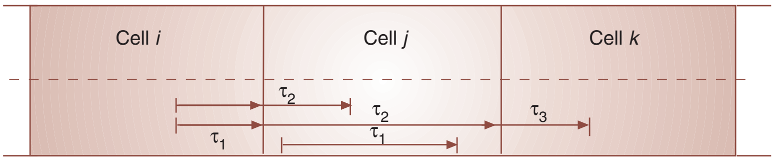


Figure 2 Dwell time and channel occupancy time

Usually the service rate is assumed to be known for the service or application under analysis, and the cross-over rate can be calculated taking into account the distribution for velocities [3]

$$\eta = \frac{1}{\int_0^{V_{max}} \left(\frac{2R}{v}\right) \cdot f(v) dv} \tag{2.3}$$

where *v* is the velocity and *f* is the velocity probability density function (note that in a linear geometry the total length of the cells is 2*R*).

For a properly designed system, the new calls traffic density increases as the coverage distance decreases, owing to the increase of the handover rate (mean number of handovers per call when the probability of the handover failure is negligible) [4]; this also causes the increase of handover failure and call dropping probabilities. The desired maximization of the new calls traffic linear density obeys to requirements of system quality, which consist of values lower than 1 – 2 % [6] for the blocking probability and lower than 0.1 – 0.5% for the call dropping probability [7]. An improvement in system performance can be achieved if guard channels are used for handover, but different solutions are obtained depending on mobility scenarios and on the number of guard channels for handover, *g*: a total number of channels *m* = *g* + *c* is considered, where *c* is the number of channels to support both new and handover calls (Figure 3).

The call dropping probability P_d is given by [4]

$$P_d = P_h P_{hf} \sum_{i=0}^{\bullet} P_h^i \left(1 - P_{hf}\right)^i \tag{2.4}$$

where *i* denotes the order of the handover and P_{hf} is the handover failure probability. For small values of P_{hf}, it can be approximated by

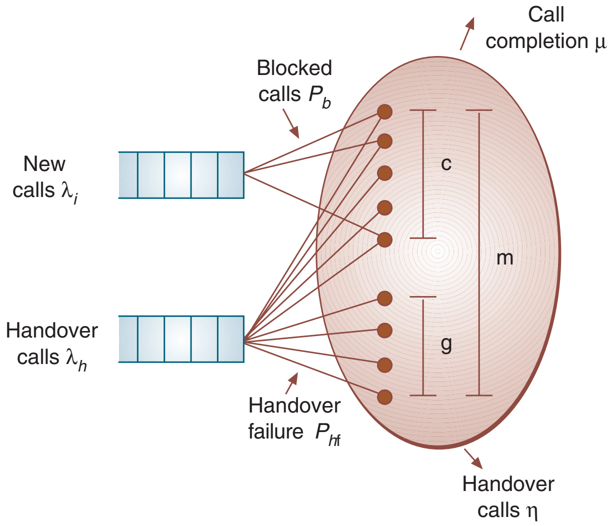


Figure 3 Model for the traffic in the case where the use of guard channels for handover is considered (extracted from [3])

$$P_d = \frac{\eta}{\mu} \cdot P_{hf} = \zeta \cdot P_{hf} \tag{2.5}$$

where ζ is the handover rate.

If guard channels for handover were not used, P_{hf} would be equal to the blocking probability P_b [10], which imposes a strong limitation because P_b would be as low as P_d determines. The use of guard channels for handover allows overtaking this limitation because P_b and P_{hf} will be decoupled [10]. In this case, depending on the coverage distance, the design is made by considering the traffic supported by m channels, from which g are guard channels [4].

The parameters involved in the design depend on the call generation rate λ, the number of channels at each cell m, besides v and μ. The simple situation of homogeneous traffic (constant value of new calls traffic in the whole service area) and linear coverage geometry (where mobiles handover between the first and the last cells, typical for circular geometry [5]) will be considered here as a first step to a more complicated (and closer to reality) analysis.

2.2 Mobility Scenarios

The scenarios examined in the analysis are presented in Table 1, where a triangular distribution, with average $V_{av} = (V_{\max} + V_{\min}) / 2$ and deviation $\Delta = (V_{\max} - V_{\min}) / 2$, is considered for the velocity [3] (Figure 4).

The probability density function is given by

$$f(v) = \begin{cases} \frac{1}{\Delta^2} \cdot [v - (V_{av} - \Delta)], & V_{av} - \Delta \le v \le V_{av} \\ -\frac{1}{\Delta^2} \cdot [v - (V_{av} + \Delta)], & V_{av} \le v \le V_{av} + \Delta \\ 0, & \text{otherwise} \end{cases} \tag{2.6}$$

which leads, when $V_{\min}, \Delta > 0$, to the following cross-over rate

$$\eta = \left\{ \frac{2R}{\Delta^2} \left[(V_{av} + \Delta) \cdot \ln \left(\frac{V_{av} + \Delta}{V_{av}} \right) - (V_{av} - \Delta) \cdot \ln \left(\frac{V_{av}}{V_{av} - \Delta} \right) \right] \right\}^{-1} \tag{2.7}$$

and when $V_{\min} = 0$ ($\Delta = V_{av}$), to the limit

$$\eta = \frac{V_{av}}{2 \cdot \ln(2)} \cdot \frac{1}{(2R)}. \tag{2.8}$$

Table 1 Scenarios of mobility characteristics

Scenario	$V_{av}[m \cdot s^{-1}]$	$\Delta\ [m \cdot s^{-1}]$
Static	0	0
Pedestrian	1	1
Urban	10	10
Main roads	15	15
Highways	22.5	12.5

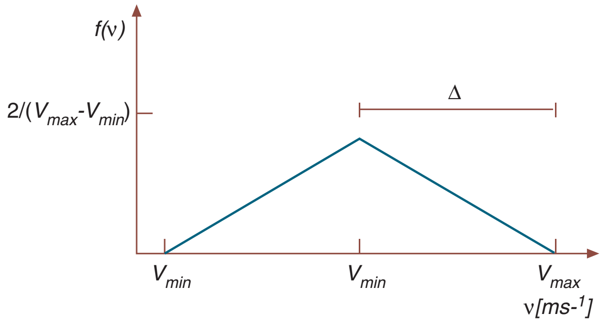


Figure 4 Velocity probability density function

Defining the average cross-over velocity η^* as

$$\eta^* = \eta \cdot (2R) \tag{2.9}$$

(η normalized to the cell length $2R$) one obtains, for the scenarios from Table 1, the values of Table 2.

The interest in defining this parameter is that it enables us to make explicit the dependence on R of some parameters to be defined later.

2.3 New Calls Traffic Linear Density without Guard Channels for Handover

For given values of m and P_b and a configuration which does not use guard channels for the handover calls, one can calculate the corresponding traffic $\rho_m = \lambda / \mu$ by using the well known Erlang-B model [13]. The new calls traffic, ρ_n, and the traffic coming from handover, ρ_h, can then be obtained as [3]

$$\rho_n = \frac{\mu}{\eta + \mu} \cdot \rho_m = \frac{2\mu R}{\eta * + 2\mu R} \cdot \rho_m \tag{2.10}$$

$$\rho_h = \frac{\eta}{\eta + \mu} \cdot \rho_m = \frac{\eta *}{\eta * + 2\mu R} \cdot \rho_m \tag{2.11}$$

where the dependence on the cell length has been made explicit by introducing the cross-over average velocity. Note that $\rho_m = \rho_n + \rho_h$.

Dividing (2.10) by ($2R$) one obtains a new calls traffic linear density

Table 2 Average cross-over velocity

Scenario	$\eta^*[m \cdot s^{-1}]$
Pedestrian	0.72
Urban	7.21
Main roads	10.82
Highways	21.21

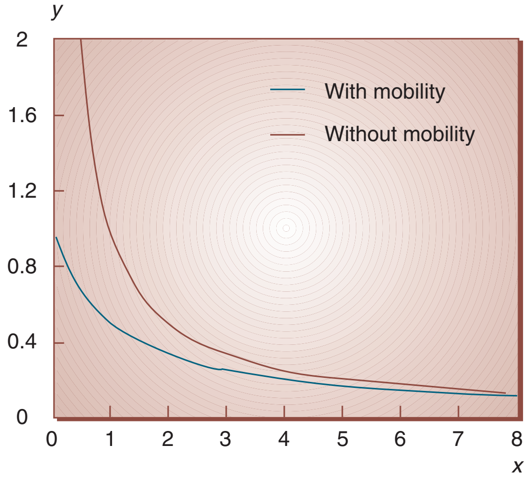


Figure 5 Normalized new calls linear traffic density y as a function of x

$$\xi_n = \frac{\rho_n}{2R} = \frac{\mu}{(\eta * + 2\mu R)} \cdot \rho_m \tag{2.12}$$

which can be normalized as follows

$$y = \frac{\xi_n}{\rho_m \cdot \mu / \eta *} = \frac{1}{1 + (\mu(2R)/\eta *)} = \frac{1}{1+x} \tag{2.13}$$

where $x = \mu(2R) / \eta^*$.

For the static scenario the new calls traffic linear density is $\xi_n = \rho_m / (2R)$ because $v = 0$ and $\eta^* = 0$. Consequently, in a non-static scenario, if the contribution of mobility was not considered, one would obtain $y = 1/x$.

From the operator's point of view the objective is to maximize ξ_n, thus the dependence of this parameter on R should be analyzed. Given constant values of the blocking probability, one obtains the graphs from Figure 5 for both situations (with and without mobility).

Then, one could conclude that ξ_n is upper limited by $(\rho_m\mu / \eta^*)$ [/m], which decreases with velocity. However, one should note that for a given R, x will be lower for higher velocities, and so y will be larger, which partly compensates for the lower values of $(\rho_m\mu / \eta^*)$ in equation (2.12).

However, call dropping probability restrictions also need to be fulfilled. Because $P_b = P_{hf}$, the blocking probability should be computed according to (2.5) [3] and, because of that, the blocking probability will be a linear function of R with slope $2\mu(P_d)_{\max} / \eta^*$.

So, the traffic supported by m channels will depend on R. For the considered scenarios an example is given in Figure 6 where: $m = 11$, $P_d = 0.5$ %, $\mu = 1/3$ min^{-1}, and the values for the average velocity V_{av} and the velocity deviation Δ are the ones presented in Table 1.

As can be seen, the supported traffic will not be constant, rather having a decreasing behaviour with the decrease of R. In this case, the dependence of ξ_n on R will be as presented in Figure 6b) for the three scenarios with higher mobility, i.e. for each scenario, there is an optimum value for the coverage distance which maximizes ξ_n. These maxima correspond to lower values of R for the scenarios with lower mobility, being lower for the scenarios with higher mobility. For coverage distances lower than this optimum value, ξ_n decreases with the decrease of R. This is due to call dropping probability restrictions, which contradicts the behaviour expected with the simple analysis without considering it. It is also worth noting that the new calls traffic linear density is much more sensitive to the mobility scenario than the supported traffic because the former, besides the dependence on ρ_m, also depends on P_h.

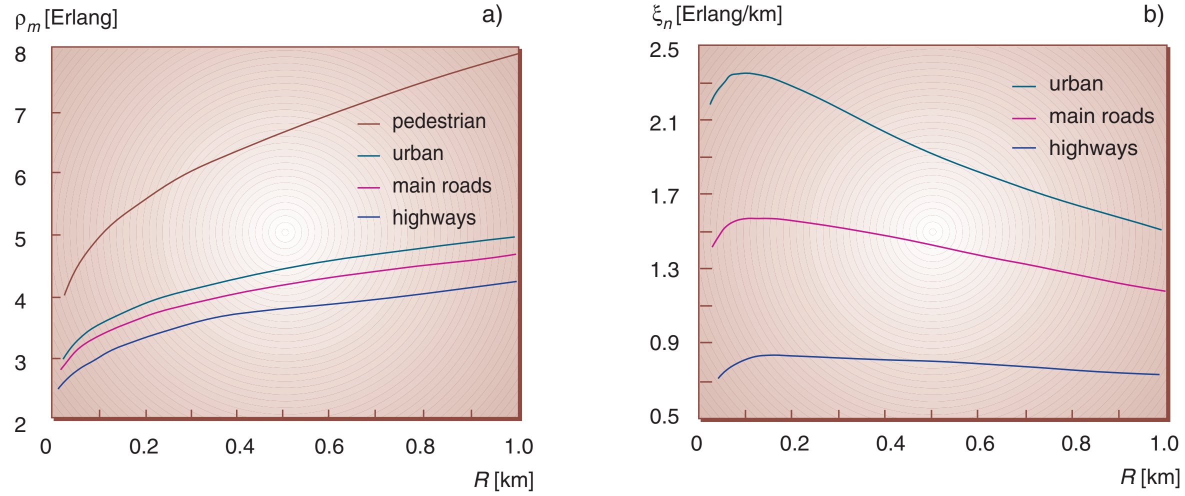


Figure 6 Traffic for m = 11 channels and the design made according to the call dropping probability restrictions a) Traffic supported, b) New calls traffic linear density

3 Traffic from Mobility with Guard Channels for Handover

3.1 Blocking and Handover Failure Probabilities

Considering the use of guard channels for handovers when no queueing of new or handover calls is performed, the blocking and handover failure probabilities are given by [3]. This is assuming that the handover traffic can be approximated by a Poisson process [8] and that the new calls traffic is also Poisson distributed, which is valid for a number of users in a cell much larger than the supported traffic.

$$P_b = \frac{(\rho_n + \rho_h)^c \sum_{k=c}^{c+g} \frac{\rho_h^k}{k!}}{\sum_{k=0}^{c-1} \frac{(\rho_n+\rho_h)^k}{k!} + (\rho_n + \rho_h)^c \sum_{k=c}^{c+g} \frac{\rho_h^k}{k!}} \tag{3.1}$$

$$P_{hf} = \frac{(\rho_n + \rho_h)^c \frac{\rho_h^g}{(c+g)!}}{\sum_{k=0}^{c-1} \frac{(\rho_n+\rho_h)^k}{k!} + (\rho_n + \rho_h)^c \sum_{k=c}^{c+g} \frac{\rho_h^k}{k!}} \tag{3.2}$$

where ρ_n is the new calls traffic and ρ_h is the handover traffic. The traffic supported by m channels is then $\rho_{m,g} = \rho_n + \rho_h$.

The new calls traffic and the traffic coming from handover can then be easily obtained as in Section 3.3, replacing ρ_m by $\rho_{m,g}$ [3]. The new calls traffic linear density is then given by

$$\xi_n = \frac{\rho_n}{2R} = \frac{1}{1 + 2\mu R/\eta*} \cdot \frac{\mu}{\eta*} \rho_{m,g} \tag{3.3}$$

For a fixed $\rho_{m,g}$ and a given μ, ξ_n is upper bounded by μ / η*, decreasing with the increase of the cross-over velocity, and having a variation with R of the type 1 / (1 + x), where x is directly proportional to R and inversely proportional to the cross-over average velocity.

3.2 Supported Traffic

The examples given here were obtained for the conditions mentioned before. For $g = 0$, using the supported traffic $\rho_{m,g}$ that verifies $P_b = 2$ %, which does not depend on R, one obtains values for the new calls traffic linear density that increase with the decrease of the coverage distance. However, the corresponding call-dropping probability constraints associated with (2.5) and with P_b being equal to P_{hf} (Figure 7) are only fulfilled in the pedestrian scenario and only for $R > 300$ m. A way to resolve this limitation without drastically decreasing the new calls traffic linear density, is the use of guard channels for handover.

From an operator's point of view, in order to achieve MBS provisional coverage distances, approximately in the range 100 – 350 m [2], one intends to increase ξ_n while R decreases. In order to obtain results for the supported traffic the procedure was the following: taking $P_d = 0.5$ %, (2.5) was used to get a value for P_{hf}. With this P_{hf} value and with $P_b = 2$ %, (3.1) and (3.2) were solved separately for the supported traffic $\rho_{m,g}$ (using (2.10) and (2.11)), and the respective values, ρ_{Pb} and ρ_{Phf}, were obtained. In order to cope with both probability requirements, the minimum of these two must be taken.

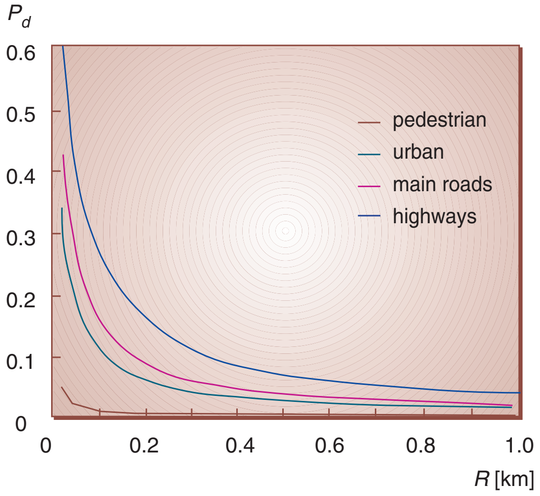


Figure 7 Call-dropping probability for g = 0

While $\rho_{Pb}(R)$ is almost constant with R (the right part of the curves in Figure 8 (example for $g = 1$)), $\rho_{Phf}(R)$ increases with R (since it was obtained according to (2.5) and η depends on R, as in (2.8) and (2.9)) (the left part of the curves). Thus, a breakpoint exists at the intersection of both curves, where $\rho_{Phf}(R) = \rho_{Pb}(R)$, and for values of R lower than this breakpoint the curves have then an appreciable slope. One can observe that the supported traffic decreases as the velocity of the associated scenario increases, mainly in the zone of the curves limited by handover failure. It can also be seen that the breakpoint occurs for increasing values of R for faster and faster mobiles.

As could be expected, when g increases, $\rho_{Pb}(R)$ decreases. However, as $\rho_{m,g}(R)$ is the minimum between $\rho_{Pb}(R)$ and $\rho_{Phf}(R)$, this decrease is only effective for the part of the curves where $\rho_{Pb}(R)$ is lower than $\rho_{Phf}(R)$. The challenge in the design for low values of R is finding values for g that, for a given m, both maximize ρ_{Phf} and keep it lower than ρ_{Pb}, mainly for the scenarios with high mobility (urban, main roads and highways).

3.3 New Calls Traffic Linear Density

For given values of m and g, we can then obtain the curves for the new calls traffic linear density $\xi_n(R)$ according to (3.3). Figures 9 and 10 show these curves for $g = 1, 2$ (the latter only for the higher mobility scenarios). For the pedestrian case, as $\rho_{m,g}$ is almost constant for all the range of R, we basically observe that it follows the behaviour of the ratio at the right member from (2.3). For the three scenarios with higher mobility, $\xi_n(R)$ presents maxima, depending on the velocity and on g. These maxima occur for distances lower than the breakpoints, for design purposes corresponding to optimum values of R, R_{opt} (Table 3), and agreeing with the provisional values for MBS and also with the need to use larger cell lengths for high mobility scenarios owing to the cost associated with signalling [12]. One can also see that the breakpoints occur for lower coverage distances as g increases.

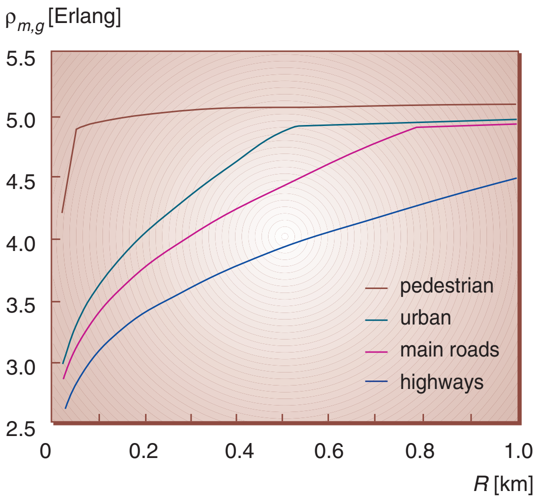


Figure 8 Traffic supported by m = 11 channels with g = 1

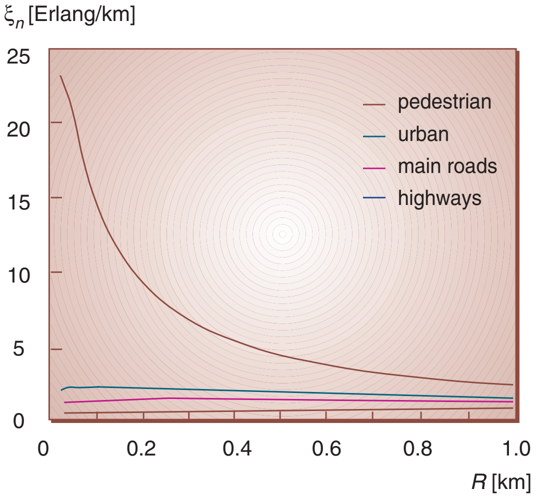


Figure 9 New calls traffic linear density, for m = 11 and g = 1

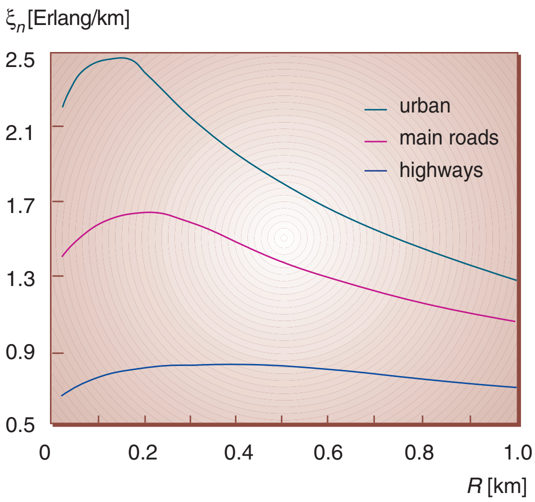


Figure 10 New calls traffic linear density, for m = 11 and g = 2

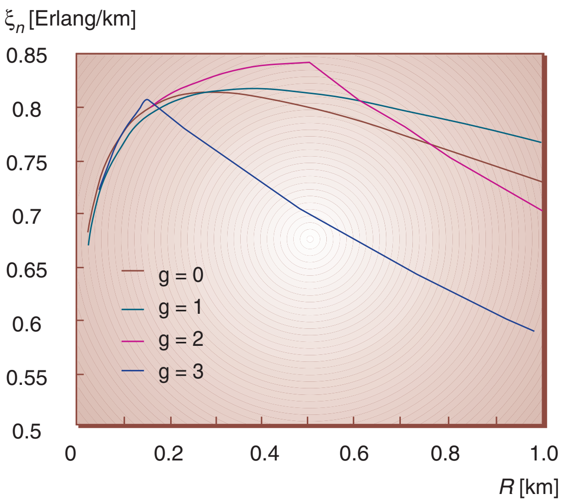


Figure 11 New calls traffic linear density for m = 11 and g = 0, 1, 2 and 3, in the highway scenario

Figure 11 presents $\xi_n(R)$ for the highway scenario for several values of g. One observes an improved new calls traffic linear density for $g = 2$ for $160 < R < 610$ m, the maximum $\xi_n = 0.84$ Erlang/km being obtained for $R_{opt} = 475$ m. An improvement of the new calls traffic linear density when $g = 3$ exists only for $130 < R < 160$ m, the maximum being obtained for $R_{opt} = 130$ m, $\xi_n = 0.81$ Erlang/km.

It is noticeable that the use of guard channels makes a difference in system performance, especially for high speed scenarios, where it allows us to overcome the problems associated with handover failure constraints. For these scenarios and for the typical coverage distances in MBS, the new calls traffic linear density is one order of magnitude below the values obtained for the pedestrian scenario (where it is approximately 10 – 15 Erlang/km), decreasing from 2.47 Erlang/km, in the urban scenario, down to 0.84 Erlang/km, in the highway scenario.

4 Conclusions

A microcellular communications system without guard channels for handover, with a linear coverage geometry, was first analyzed. Models to compute the supported traffic and the new

Table 3 Approximate values for R_{opt} and maximum values for ξ_n with m = 11

	R_{opt} [m]		ξ_n [/km]	
Scenarios	$g = 1$	$g = 2$	$g = 1$	$g = 2$
Urban	125	150	2.40	2.47
Main roads	175	250	1.60	1.65
Highways	375	475	0.82	0.84

calls traffic linear density as a function of velocity and cell length were examined. For a fixed blocking probability, one verifies that the new calls traffic linear density, which was used as a measure to system capacity, is upper limited by the average cross-over velocity of the associated scenario, having lower values for scenarios with high mobility, and varying differently depending on the average cross-over velocity. However, for actual scenarios the call-dropping probability constraints are not fulfilled in such conditions. Thus, in order to simultaneously verify the constraints for the blocking and the handover failure probabilities, it is necessary to consider lower blocking probabilities which leads to a decrease in the new calls traffic linear density with the decrease of R for such configurations. Feasible values for the new calls traffic linear density were obtained, where there are optimum values for the coverage distance that maximize it. These maxima correspond to lower values of the coverage distance for scenarios with lower mobility, being lower for scenarios with higher mobility.

In order to overtake the limitation associated with the case where no guard channels are used, configurations that use guard channels for handover were analyzed. Results were obtained for different values of the number of guard channels g. One concludes that, for the coverage distances foreseen for MBS, higher optimum values for the new calls traffic linear density are obtained for $g = 2$. However, for the highways scenario the corresponding coverage distances are larger than the foreseen cell lengths. It has also been verified that there is a degradation in system capacity, measured in Erlang/km, for higher and higher mobility scenarios.

Acknowledgements

Fruitful conversations are acknowledged with Prof. João Luis Sobrinho of IST and with Prof. Vladimir Gligic from University of Beira Interior. The authors would like to express their sincere thanks to the Department of Electromechanical Engineering of the University of Beira Interior, for providing the computational resources that enabled part of the work of Fernando Velez to be successfully accomplished.

References

1 Fernandes, L. Developing a system concept and technologies for mobile broadband communications. *IEEE Personal Communications Magazine,* 2, (1), 1995, 54–59.

2 Velez, F, Correia, L M. Optimization criteria for cellular planning of mobile broadband systems in linear and urban coverages. In: *Proc. ACTS Mobile Communications Summit, Aalborg, Denmark,* Oct. 1997, 199–205.

3 Chlebus, E, Ludwin, W. Is handoff traffic really Poissonean? In: *Proc. of ICUPC' 95 – IEEE International Conference on Universal Personal Communications, Tokyo, Japan,* Nov. 1995, 348–353.

4 Jabbari, B. Teletraffic aspects of evolving and next-generation wireless communication networks. *IEEE Personal Communications Magazine,* 3, (6), 1996, 4–9.

5 Sidi, M, Starobinski, D. New call blocking versus handoff blocking in cellular networks. *Wireless Networks,* 3, (I), 1997, 15–27.

6 Correia, L M et al. *Report on design rules for cell layout.* RACE Deliverable R2067/IST/2.2.3/DS/P/044.b1, RACE Central Office, Brussels, Belgium, 1994.

7 ITU-T. *Networks grade of service parameters and target values for circuit-switched public land mobile services.* ITU-T Recommendation E.771, Geneva, 1996.

8 Ross, K, Tsang, D. Teletraffic engineering for product-form circuit-switched networks. *Adv. Applied Probability,* 22, 1990, 657–675.

9 Guérin, R. Channel occupancy time distribution in a cellular radio system. *IEEE Transactions on Vehicular Technology,* 36, (3), 1987, 89–99.

10 Hong, D, Rappaport, S. Traffic model and performance analysis for cellular mobile radio telephone systems with prioritized and non-prioritized hand-off procedures. *IEEE Transactions on Vehicular Technology,* 35, (3), 1986, 77–92.

11 Yacoub, M D. *Foundations of mobile radio engineering.* CRC Press, Boca Raton, Florida, USA, 1993.

12 Silva, N, Brázio, J. Reduction in location management signalling by the use of subpopulation-tailored location area sizes. In: *Proc. of ICUPC' 96 – International Conference on Universal Personal Communications, Cambridge, Massachusetts,* USA, 1996, 602–606.

Fernando J. Velez (28) is Lecturer at the Dept. of Electromechanical Engineering at Univ. of Beira Interior, Covilhã, Portugal, and PhD student in Electrical and Computer Engineering (ECE) at IST, Technical University of Lisbon. He holds a degree in ECE from IST and received his MSc in ECE in 1996. His research interests include cellular planning, cost/revenue performance of mobile communication systems and traffic/mobility aspects. He is currently working on multi-service traffic in mobile broadband systems.

e-mail: fjv@ubi.pt

Luis M. Correia received his PhD in 1991, and is currently a Professor at IST in the area of Telecommunications. His research work is focused on wave propagation, antenna design, traffic modelling and cellular planning in mobile communications. He has been involved in European research, currently in ACTS, as well as Chairman in COST 259. He has also served as Auditor/Evaluator in the ACTS and ESPRIT programmes.

e-mail: luis.correia@lx.it.pt

The *haves* and *have-nots* – The distribution of ICTs in Norwegian households

RICH LING

- **The interactive communication technologies (ICT) *haves* – defined here as those who consume far more than the average number of electronic devices – are dominated by those in the younger age groups while the *have-nots* are largely over 55. Analysis shows that the group of elderly *have-nots* is becoming smaller and smaller as the younger groups carry their consumption practices with them as they age. There seems to be a shift in the consumption of technology between those born before and after the Second World War.**
- **Income is perhaps the single most important factor determining one's status as a *have* or a *have-not.***
- **Attitudinal factors and one's 'ideological' orientation towards ICT are important in determining one's consumption of these devices.**
- **The greatest attitudinal strife between the *haves* and the *have-nots* is found among mature adults aged 35 – 55. The *haves* and the *have-nots* in this age group disagree about the role of ICT in daily life, the ability to integrate it into the home and in their general perceptions of systems such as the Internet.**

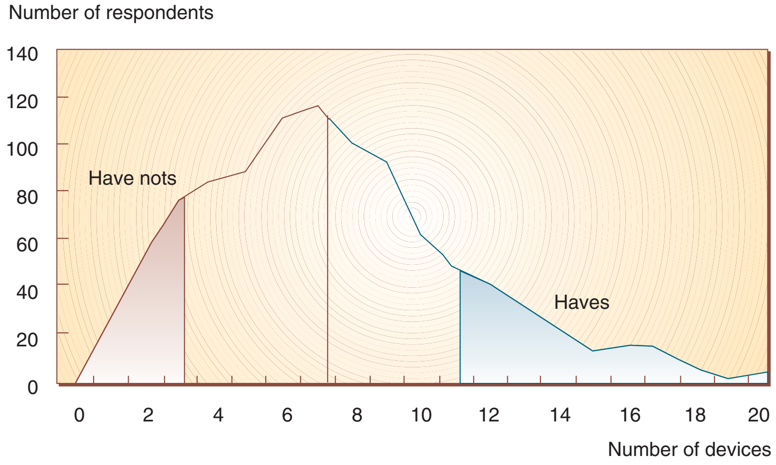


Figure 1 Distribution of electronic devices among the Norwegian population

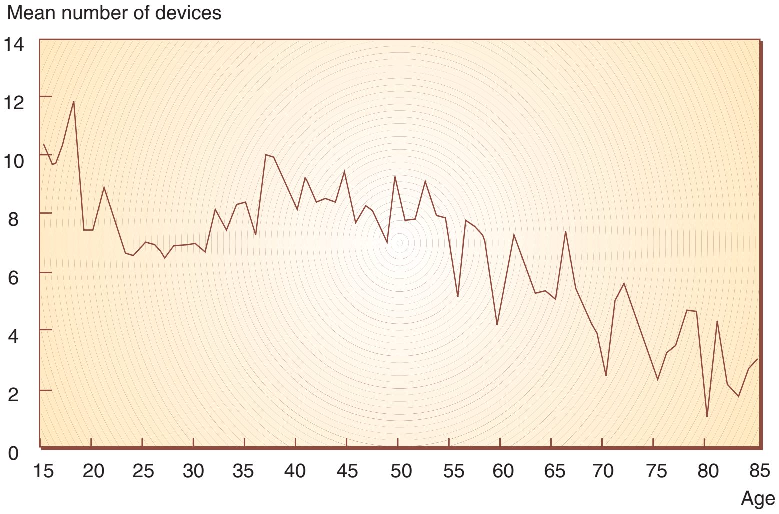


Figure 2 Mean number of devices

Introduction

A great deal of interest has been directed towards the question of the penetration of ICTs into the private sphere. There is a concern that we are developing a dichotomous society of those who have and those who do not have these devices. The affluence of Norway has led portions of the population to become quite advanced consumers of electronic devices while others lag behind in this activity. In this paper we focus on two social groups with quite opposite life experiences, namely the ICT *haves* and *have-nots* in Norway. As one will see in this paper differences in one's consumption of ICTs reveal significant differences in age, education, income, the use of media along with differences in attitude on a range of issues.

In Figure 1 we have displayed the distribution of devices from one of the two databases used to analyse this phenomenon[1]. It is the cases at the high and the low end of this distribution that are defined as the *haves* and the *have-nots*[2].

Age makes a big difference

One of the first things that strikes the reader is variation in consumption of electronic devices by age. In Figure 2 we see that the mean number of devices in the home generally declines as the age of the interviewee increases. There are, however, nuances in this picture. One can see that interviewees between 18 and 35 reported fewer devices than the youngest respondents. This represents the fact that they had recently moved out of the parents' home, had a less stable economy and a more nomadic existence. Respondents between 35 and 55 years of age had a relatively large number of devices while, as we will see below, those over 55 reported the ownership of fewer and fewer devices.

[1] *The 'Telenor R&D database' analysis included the following devices in the inventory of ICTs: PC, printer, modem, CD-ROM, more than one TV, cable, satellite dish, video, CD player, game console, telefax, copy machine, mobile telephone, more than one telephone handset, more than one telephone line, telephone answering machine, video camera. The Statistics Norway analysis included the following inventory of items: PC, CD-ROM, modem, Internet, CD player, one or more TVs, video player, cassette player, record player. Finally, the analysis of generational effects is based on the following inventory of items: one or more TVs, video player, game console, PC.*

[2] *The distribution has the general form of a normal curve. The mean is slightly higher than 7 items and the first standard deviation is 3.7 units above and below the mean. The groups are roughly the same size but since the definition units are discrete devices it was impossible to establish two equally sized groups.*

Figure 3 shows the distribution of the *haves*, *have-nots* and the remaining population by age[3]. The distribution is divided into five age-based groups. These include *youth* (younger than 18 years of age), *young adults* (18 – 35), *established adults* (35 – 55), *'pre-retirement' adults* (55 – 65), and *retirees* (older than 65). The *youth* and the *retirees* are groups defined by their relation to age-bounded social institutions, namely the school system and work life. The figure shows a rapid drop in the percent of *haves* as this group migrates into the *young adult* phase.

The remaining three groups are defined empirically. The *young adults* are in the process of establishing families and the pursuit of careers. The relatively small number of haves in this group points to a certain economic pressure here. The percent of *haves* was higher for respondents between 35 and 55. This is a consequence of the fact that the *established adults* have generally overcome the first hurdles of establishing themselves and enjoy a greater stability than the previous group. The *established adults* have the income with which to purchase the various ICTs and they are also often pressured to purchase by teenage children living at home.

As the respondents moved into their mid fifties, i.e. the *'pre-retirement adults'* group and the *retirees*, we see that the percent of *have-nots* rises rapidly. This trend continues through the remainder of the distribution. These two groups seem to have a different relationship to ICT than those who are younger. Several elements may conspire to form this situation. Many work and home related ICTs have arrived on the scene only after this group was well established in their careers. Accommodating them in their everyday routines may have been more difficult than for those who were introduced to the technologies at a younger age. These two groups do not have the same domestic pressure to participate in the ICT revolution as teenage children have generally moved out of the home. Finally, the *youth*, *young adults* and the *established adults* were born after the Second World War. The *'pre-retirement adults'* group and the *retirees* were born before this dramatic historical event. While the former grew up during an extended period of prosperity and belief in technology, the latter grew up during the end of the depression and the hardships of the war. This early life experience may have had an effect on their willingness to use money on relatively expensive devices.

This argumentation suggests that as the *established adults* migrate into the older age groups they will carry with them their taste for ICTs. To examine this we have compared the percent of *have-nots* using a standard list of ICT holdings between 1994 and 1997. The data shown in Figure 4 indicates that the high percentage of *have-nots* associated with the elderly is being pushed into older and older age groups. The data shows, for example, that while in 1994 the point at which 20 % of the elderly were *have-nots* was at age 50, this had risen to 56 by 1997.

This data indicates that one can expect the elderly *have-nots* to become a smaller and smaller portion of society as those with ICT experience move into this age group. However, as we will see below, limited income in this group along with the introduction of ever new technologies such as Digital Video Display (DVD) and Web TV will moderate the reduction of *have-nots* among the elderly.

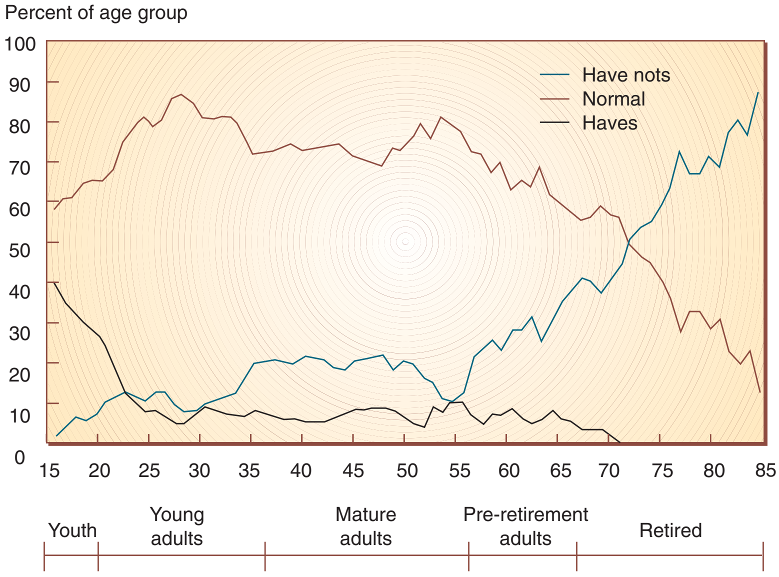
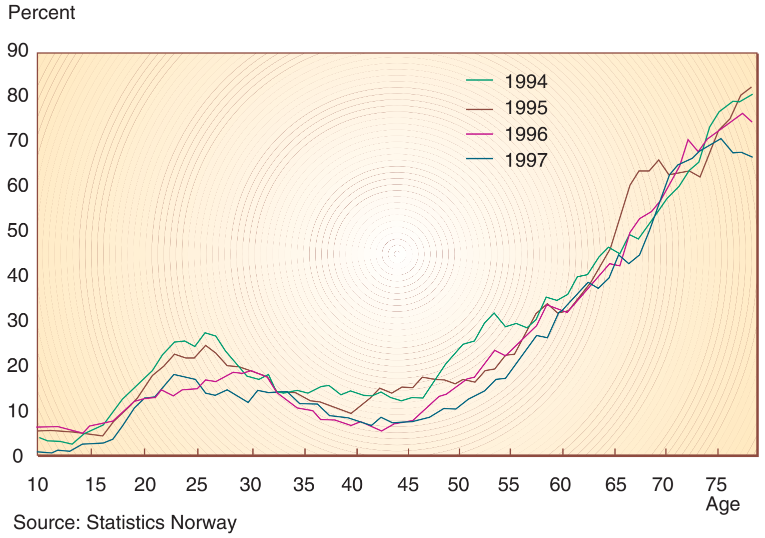


Figure 3 Percent of age group



Source: Statistics Norway

Figure 4 Aging of the have nots

Income is a key factor

Changing gears now, we will examine the five groups described above for differences in income, education and media use. The data dictates, however, that in the following analysis we must

[3] *The data in Figures 3 and 4 is represented by a seven-year moving average in order to flatten out the more extreme swings in the data.*

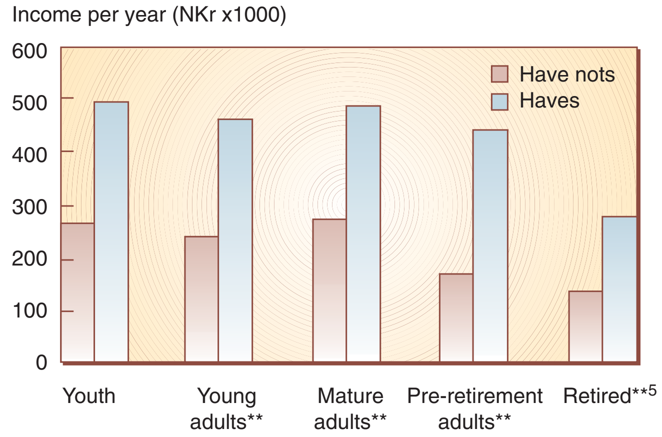


Figure 5 Income per year[5]

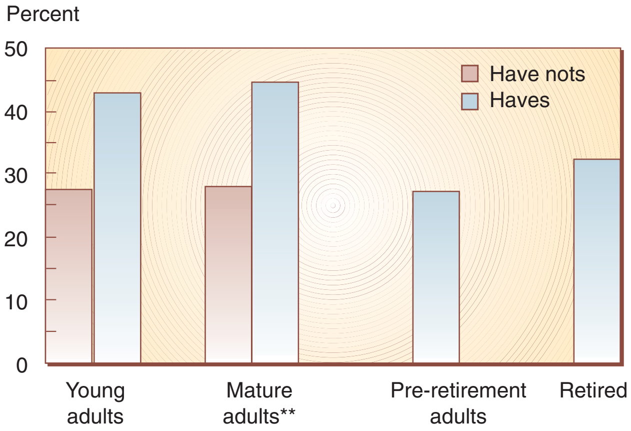


Figure 6 Percent of respondents with higher education

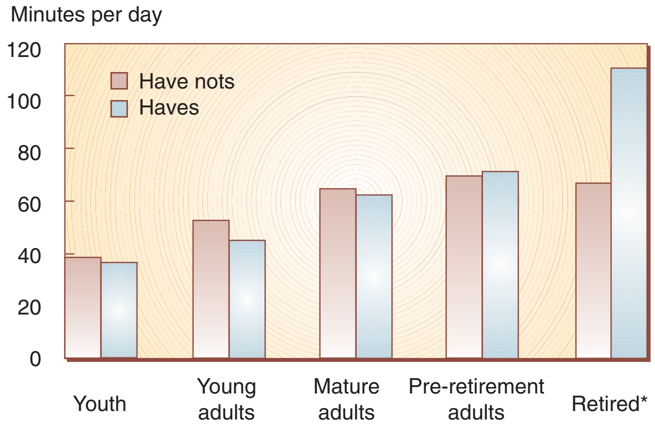


Figure 7 Time per day spent on reading newspapers

define the *haves* and the *have-nots* age-group by age-group as opposed to the a single definition across all groups as has been used in the analysis up to now[4].

The analysis indicates that a major factor determining one's status as a *have* or a *have-not* is the household's income. Figure 5 shows that there were significant income differences within all five age groups. Even though the age differences were slightly more pronounced in the lower age groups, they were quite considerable in every case. It is often the case that gender, educational/career choice and income are related factors as women generally have lower incomes. In our analysis, however, we examine household income that tends to disguise this difference.

Another variable that showed widespread differences between the *haves* and the *have-nots* was education. In Figure 6 we see the percent of respondents who had completed a college education in each of the age groups[6]. The data shows that there were strongly significant differences between the *haves* and the *have-nots* in the last three groups. Our data, in fact, showed that it was difficult to find *have-nots* with a college education.

Another aspect of the differences between the *haves* and the *have-nots* is the use of other media. The data here shows that the *haves* among the *retirees* are oriented towards reading while the *haves* among the younger age groups are higher than average telephone users.

In Figure 7 we see the mean number of minutes per day reported for reading newspapers. The interesting point here is the significant difference between the *haves* and the *have-nots* among the retirees. While all other age groups were quite similar on this point, there was a considerable difference in this age group. This difference is likely a consequence of the educational differences pointed out above.

At the other end of the age spectrum we find a difference in the number of telephone calls made per day by the *haves* and the *have-nots*. While there were differences across all age groups here, it was among the *youth* and the *young adults* that we found the significant differences between groups.

Attitudinal differences: Established adults show the greatest differences

In the data we also gathered information on the respondents' attitudes towards ICT. Our analysis shows that these were as strong an indicator of one's *have/have-not* status as the more traditional background variables examined above. Thus, we

[4] *In practical terms, the use of a single definition across all ages would mean that there are too few* haves *among the retirees just as there are too few* have-nots *among the youth. Thus,* haves *and* have-nots *were defined as those above and below the respective standard deviation for each age group.*

[5] *The following symbols are used in the remaining figures: + sig. = 0.062, * sig. < 0.05, ** sig < 0.001.*

[6] *The youngest age group was excluded from this analysis as they were still in the educational system.*

were able to not only point out demographic differences between the groups, but also examine some of the attitudinal differences between the groups. Specifically we included a series of 11 questions in the questionnaire examining their attitudes toward the Internet and towards PCs at home. The questions covered various aspects of ICT such as the perceived importance of the Internet in future work situations, the degree to which the PC belongs in the home and general glosses of the Internet.

The first point to be made in this analysis is that with the exception of the *established adults* there is little disagreement between the *haves* and the *have-nots* in regards to the role of the Internet. By contrast, the data shows a complex of differences between the *haves* and the *have-nots* who are over 35 years of age with regard to the PC.

Somewhat tellingly, *have nots* among the *retirees* felt that the PC was too complex for use in the home and that it represented a disturbing element in the home. In Figures 9 and 10 one can see the relative weight that each of the age groups placed on these items on a five-point scale. Figure 9 in particular shows a dramatic increase in the degree to which the *have-nots* among the *retirees* felt that the PC was too complex for use in the home. Figure 10 points to the degree that the respondents felt that the PC was a disturbing element in the home. In this case there were significant differences between the *haves* and the *have-nots* for both the *retirees* and the *established adults*.

The attitudinal data indicates that the ideological battleground over ICT in the home is among the *established adults*. It is in this age group that one finds the most extensive differences between the *haves* and the *have-nots*. There was disagreement as to the extent of ICT's impact in society. This can be seen in the disagreement over the statements "Eventually everybody will buy a PC", "Knowledge of how to use a PC will be important in school and work", "The Internet has importance for working life" and "The Internet is a fad". In addition, there was disagreement as to the ability to integrate PCs into the home "The PC can disturb domestic life" as well as the general understanding of the Internet "The Internet is only porn and violence". Figure 11 shows the relative disagreement between the *haves* and the *have-nots* on these items for the *established adults*.

Conclusion

In conclusion, our analysis has shown that the *haves* are dominated by those in their teens and those in their 30s, 40s and early 50s. The *have-nots* are largely over 55. Analysis shows that the elderly *have-nots* are being replaced by younger cohorts who are bringing their consumption patterns with them into the *'pre-retirement adults'* and *retirement* phases of life. The analysis also shows that the orientation of those born before and after the Second World War seems to be dramatically different in regard to the consumption of ICT. While the former are comfortable with the consumption of a variety of items the latter are more grudging in their use of these devices.

Income and education are important clarifying factors when understanding one's interest in the consumption of ICT. The higher the income and the higher the educational level the more likely one is to be among the *haves*. Educational level is a particularly important factor for the older respondents.

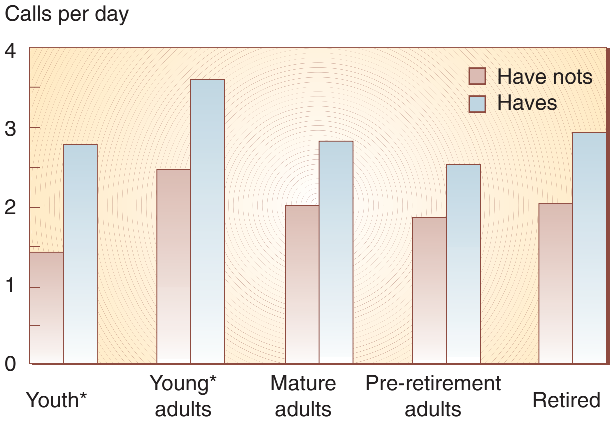


Figure 8 Use of telephones

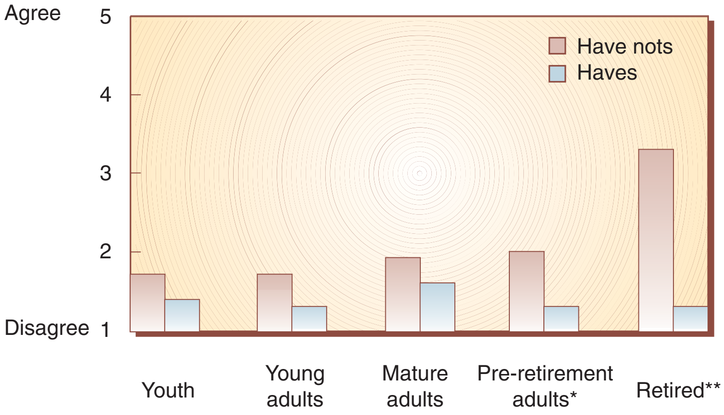


Figure 9 The PC is too complex for use in the home

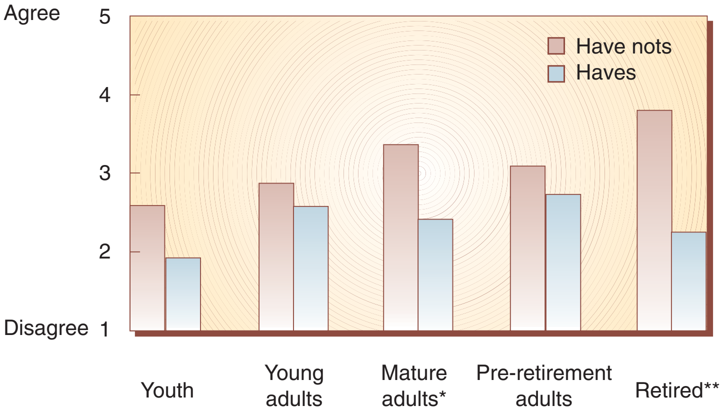


Figure 10 The PC is a disturbing element in the home

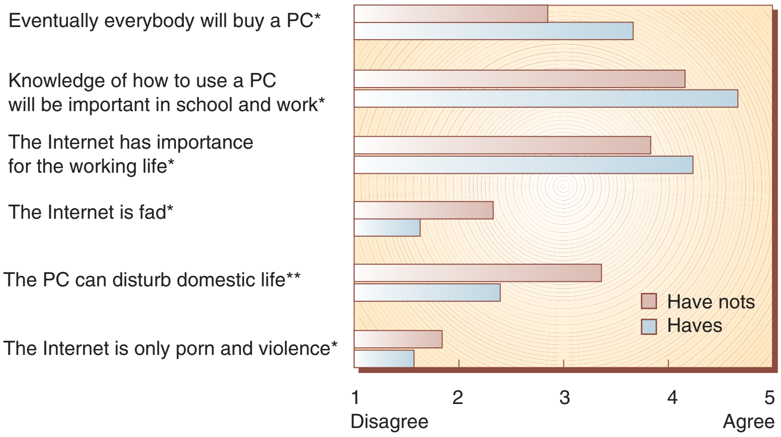


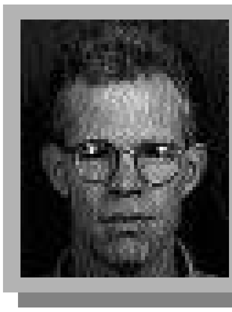
Figure 11 Disagreement among established adults

Finally we find that one's 'ideological' orientation towards ICT is important in determining one's consumption of these devices. With the exception of the *established adults* aged 35 – 55, the ideological orientation towards the PC is not problematic. Either the age group is uniformly comfortable or uniformly uncomfortable with this technology. By contrast, the age groups over 35 are far more divided in their attitude towards the Internet.

When looking at the specific age groups, the data shows an ideological divide between the *haves* and the *have-nots* among the established adults aged 35 – 55. Here one sees disagreement as to the role of ICT in daily life, its integration into the home and in the general perception of ICT services.

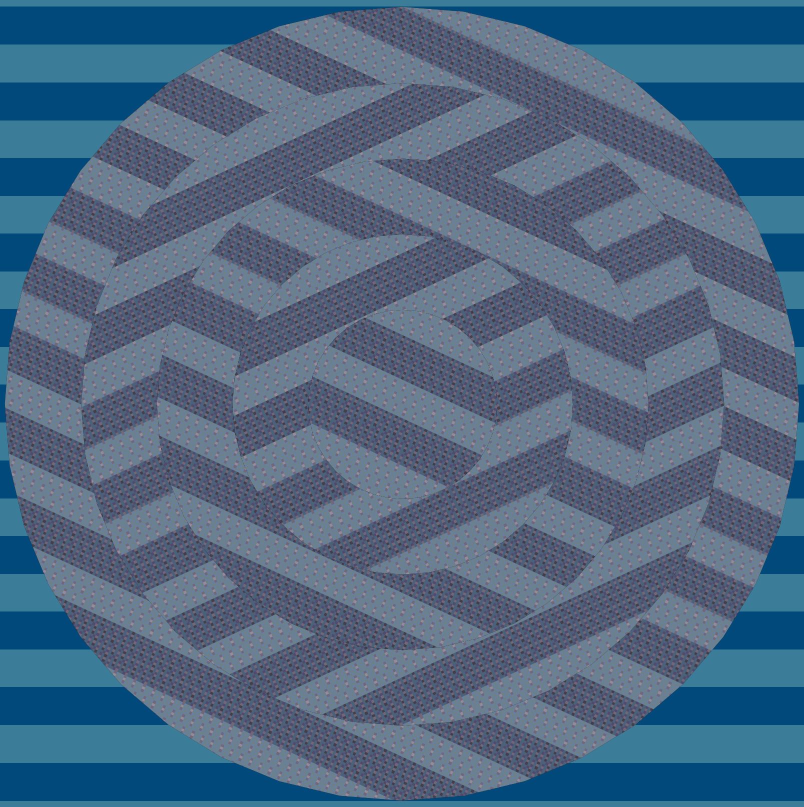
Method

The data for this paper comes from two main sources. The main analysis focuses on Statistisk sentralbyrå's (Statistics Norway) *Mediebruksundersøkelse* from 1997 and from data gathered by Telenor R&D in 1997 on PC ownership. The former source included 2000 interviews carried out at four points spread throughout the year. The population of this survey was a representative sample of persons aged 9 to 79. The latter includes 1000 interviews carried out in the spring of 1997 wherein the respondents were a representative sample of Norwegians ranging from 13 to 91 years old. Both questionnaires gathered information on the consumption of domestic electronic devices. The R&D survey had a somewhat more extensive list of devices than did the Statistics Norway survey. The R&D survey included items such as TV, radio, video machine, PC, CD player, fax machine, copier, telephone answering machine and the like.



Rich Ling (44) is a sociologist working at Telenor R&D, Kjeller. He received his Ph.D. in sociology from the University of Colorado, Boulder. Since coming to Norway he has worked at the Resource Study Group founded by Jergan Randers and has been a partner in the consulting firm Ressurskonsult. For the past five years he has worked with Telenor and has been active in researching issues associated with new information technology and society.
e-mail: richard.ling@fou.telenor.no

Kaleidoscope



NORDTEL
– 80 years of Nordic co-operation in the telecommunications area

EINAR UTVIK AND VILLY TOFT

1917 – 1997 in chronological perspective

The telecommunications administrations of Sweden, Denmark and Norway started co-operation on an occasional basis as well as bilateral discussions already in 1858, but the more permanent Scandinavian, and some years later, the genuinely Nordic telecommunications co-operation, started in 1917. During World War I the governments of the three countries had started a co-operation on foreign policy based on their common interests as neutral states. The telecommunications administrations were also involved in this co-operation, with a particular view to making the censorship of the telegraph and telephone traffic more effective. Such censorship was considered necessary in order to prevent the transmission of information that could endanger the security of Scandinavian merchant vessels, and it was also desirable to keep the propaganda of the belligerent powers within reasonable limits.

During such a government conference in Copenhagen in December 1917, the directors general of the three telecommunications administrations took the opportunity to conduct internal discussions on matters purely related to telecommunications, and this meeting has later been considered as the first Scandinavian telecommunications conference. Such discussions were pursued on a regular basis, and from the eighth meeting in Stockholm in October/November 1924, there was also participation by the Finnish and Icelandic telecommunications administrations. Consequently, from this conference onwards, the co-operation can be considered as being genuinely Nordic. The reason why Finland and Iceland were invited to this particular conference was that the agenda included discussion of proposals that were to be presented to the International Telegraph Conference in Paris in 1925.

The Nordic telecommunications conferences (NT) were held with irregular intervals until the twenty-first meeting in Helsinki in December 1947. Subsequently, the conferences took place regularly every second year, with the exception of the extraordinary conferences in 1954 and 1964.

The conference venues alternated between the different countries in such a way that the meetings were held in Iceland in the years x1, in Denmark in the years x3, in Sweden in the years x5, in Finland in the years x7 and in Norway in the years x9. The duration of the conferences was normally 4 – 5 working days.

To a certain extent, visits to industrial undertakings or sightseeing tours were organized in between the regular conference sessions. In addition to trade-related and cultural inspiration, these supplementary activities offered possibilities for more informal, collegiate discussions between the participants.

During the telecommunications conferences, the questions on the agenda normally covered points handled by the three standing committees: the Traffic and Telecommunications Committee, the Technical Committee and the Radiocommunications Committee.

At the thirty-fourth meeting of NT in Kabelvåg, Norway, in June 1969, the Swedish delegation proposed an extension of the Nordic telecommunications co-operation. The main reason for this proposal was the substantial increase in the number of items on the agenda of the conferences, and it was considered necessary that the activities of the conferences were treated in a more systematic and long-sighted perspective, i.a. by establishing permanent groups/committees to work more continuously in the periods between the telecommunications conferences. To study this proposal more in detail, the meeting established an expert group under the chairmanship of Torsten Larsson, Sweden. The task of this group was to work out a detailed proposal for the organisation and form of the further co-operation. The final proposal of this expert group was presented to the thirty-fifth meeting of NT, in Reykjavik in 1971.

The debate on this proposal at the conference resulted in the signing of the 'General agreement on Nordic co-operation in the telecommunications area' in Reykjavik on 27 August 1971. As provided in this agreement, the Nordic Telecommunications Conferences (NT) should continue to be held every second year alternating between the Nordic countries and pursuant to the same rules as those previously applied.

To supervise the NT activities between the conferences, a steering committee, NST, was established, consisting of the five directors general and under the chairmanship of the director who was to host the next NT conference. Three co-ordination committees were also created to assist the NST in the management of the different working groups:

- NTD – for operational and tariff questions
- NTR – for technical questions related to radio communications
- NTT – for technical questions related to telecommunications (excluding radio communications).

At the thirty-sixth NT conference in Rold Storkro, Denmark, in June 1973, three additional committees were established:

FOTO: BESSING FOTO

The cable ship "Peter Faber" with submarine cable plough

- NTSK – Nordic Telecommunications Satellite Committee
- NKTK – Nordic Cable TV Committee
- TS – Tanum Board (for the operation of the common satellite earth station in Tanum, Sweden).

Seventeen working groups presented their reports to this conference, but already four years later, in 1977, the number of working groups had increased to thirty.

At that time – during the thirty-eighth NT conference in Savonlinna, Finland – the Danish delegation proposed that the rules for the future Nordic telecommunications co-operation should be revised. The principal reason for this was that even if important results had been achieved, it was a fact that the co-operation required great resources, and the main idea behind the proposal was to rationalise the co-operation activities through a partial delegation of the management responsibility of NST to the co-ordination committees, i.a. by giving these the competence of establishing and disbanding working groups. In this way, NST would have more time to treat fundamental and principal, future-oriented questions and problems.

NT decided to create a working group to handle this matter and to present a proposal for a new agreement at the next NT conference.

The new agreement, 'Agreement and guidelines for Nordic co-operation in the telecommunications area', was signed at the thirty-ninth NT conference in Stavanger, Norway, on 31 May 1979.

In this agreement it was stated that "the permanently organised Nordic co-operation takes place in Nordic telecommunications conferences (NORDTEL), in committees and working groups".

The following committees were created to supervise the working groups within their respective areas of competence:

- NM – the committee for marketing, network planning and tariff questions
- NT – the committee for technical questions related to telecommunications (excluding radio communications)
- NR – the committee for technical questions related to radio communications. This committee was also the supervisory

FOTO: NN

NORDTEL meeting, Faroe Islands, June 1993. Walking on the beach after inauguration of the CANTAT-3 terminal station at Tjörnavik body of NTSK (Nordic telecommunications satellite committee)

- ND – the committee for data communication
- TS – the Tanum Board.

This agreement provided a well suited framework for the activities during the 10-year period from 1979 to 1989.

At the end of this decade, however, it became clear that the gradual liberalisation and growing competition in the telecommunications area in Europe made it increasingly difficult for the working groups to collaborate closely on commercial and also on technical questions. It was therefore considered necessary to revise the current agreement, and after long and thorough preparations in a working group, a new agreement was signed at the NORDTEL meeting in Tórshavn on 8 June 1993.

This new agreement implied the disbanding of the former NM, NT and NR committees (ND had already been disbanded some time earlier).

Two new groups were established under NORDTEL: NS, the committee for network strategy, and KM, the contact group for market development. There was also agreement on maintaining NP, the committee for network planning, which had been split off from NM in 1989 as a separate committee.

About four months earlier, on 17 February 1993 in Copenhagen, the five managing directors of the NORDTEL member organisations had also established a new trade association for telecommunications operators in the Nordic countries, NTOB, 'Nordic Telecommunications Operators' Trade Association'. The association was open for membership by all telecommunications network operators in the Nordic market and provided a separate legal identity for the trade co-operation (association registered in Copenhagen, subject to relevant Danish legislation).

According to its statutes, the association should also have a separate office in Brussels, Belgium. This office was established in August 1993, at the same time as the first director of NTOB, Mr Svein Tenningås, Televerket, Norway, took up his duties which mainly consisted of discharging NTOB's functions in Brussels.

Within NORDTEL, it proved increasingly difficult to continue the activities within KM, the contact group for market development, because of the steadily growing competition. This called for another revision of the NORDTEL agreement, and the revised agreement was signed at the NORDTEL meeting in Gothenburg, Sweden, on 8 August 1995. The new agreement implied the disbanding of KM while the two other committees

- NS – the committee for network strategy, and
- NP – the committee for network planning

continued their activities.

After the establishment of the trade association, NTOB, the majority of the members set up their own, individual office in Brussels. Taking into account the fact that in addition to Denmark, Sweden and Finland had become members of the European Union, the NTOB board decided in July 1996 to close down the association's office in Brussels. The office was closed on 1 September 1996, the date of termination of the director's contract with the association. At the same time the board decided to consolidate the association's secretarial activities in Copenhagen and to appoint Mr Villy Toft, Tele Danmark, as director of NTOB for a period of nine months.

FOTO: ULF OWENEDE

NORDTEL meeting, Gothenburg, August 1995

During these nine months a revaluation of NTOB's tasks and organisation was to take place. In this revaluation, the stated wish of NORDTEL to consider the possibility of merging the two organisations and their activities should be taken into account.

As a result of this revaluation, the NORDTEL members (the managing directors of Telia, Telenor, Tele Danmark, Telecom Finland and P&T Iceland) agreed at their meeting in Porvoo, Finland, on 13 May 1997, to merge NORDTEL with the four year old trade association, NTOB, with a clear recommendation of using the name NORDTEL, and not NTOB, for the merged organisation.

In conformity with this decision, the general assembly of NTOB agreed to the proposed merger at its meeting in Copenhagen on 28 May 1997 and amended the statutes to provide the formal basis for the new, merged organisation. The subsequent meeting of the NTOB board, which was held the same day, decided to accept Tele Danmark's offer to make Villy Toft available for the position as NTOB director for another year.

As a final consequence of this reorganisation of the Nordic telecommunications co-operation, the members of the former NORDTEL organisation, at their meeting in Oslo on 4 December 1997, decided to terminate the agreement signed in Gothenburg in August 1995 and to abolish the old organisation.

Overall survey

Initially, the Nordic collaboration was characterised by questions relating to tariff matters. The main purpose of these efforts was to keep tariffs for inter-Nordic telecommunication traffic as low as possible without disregarding requirements for a satisfactory economical result of the telecommunication activities.

After World War II the collaboration was extended to include joint economical activities as well. A number of successful and positive achievements emanating from this collaboration could be mentioned, in particular from the 1970s and 1980s when the co-operative activities were at their peak:

- Co-operation on joint planning, extension and exploitation of inter-Nordic traffic connections and connections from the Nordic countries to other countries, often organised as a pool co-operation. This co-operation resulted in significant economical savings for all the Nordic telecom operators.
- Joint development and establishment of NMT – the Nordic Mobile Telephone system – which turned out to be a great success and which laid the foundation of the outstanding position of the Nordic countries within mobile telecommunications.
- Joint development and establishment of the public Nordic data network (DATEX).
- Joint establishment and operation of the satellite earth stations in Tanum, Eik and Ågesta for the INTELSAT, INMARSAT and EUTELSAT systems respectively. The first satellite earth stations required so huge investments that it would have been very difficult for the individual Nordic telecom operator (administration) solely to meet such expenses. By taking this up as a joint activity all the Nordic telecom operators (administrations) became involved in the satellite communications at an early stage.
- Collaboration on joint representation in the satellite organisation INTELSAT. Each of the individual Nordic telecom operators (administrations) was too small to obtain a seat on INTELSAT's Board of Governors, but by forming a group, a seat in the Board of Governors and in the sub-committees was ensured.
- Co-operation within European and international organisations such as ITU, CEPT, ETSI, EURESCOM, ETNO etc. This co-operation was established very early (ref. joint preparations for the International Telegraph Conference in Paris, 1925) and has since characterised and strengthened the participation of the Nordic telecom operators (administrations) in the international telecommunications organisations. Today this co-operation constitutes an important part of the activities in the new Nordic trade association.

Summing up, the tradition of a constructive co-operation among the major incumbent Nordic telecommunication operators has lasted for nearly a century.

Despite great differences between the Nordic countries in topography, population density and differences in the national organisation of the telecom enterprises, common cultural, political

FOTO: BILDCENTER - FARSTA

NORDTEL (NTOB) Board meeting, Stockholm, October 1997 and demographic characteristics have formed the basis of significant mutual interest also within the telecommunication area.

A number of the achievements have been quite tangible (ref. the examples mentioned above), whereas others have had their impact through strategic political perspectives and may therefore have been less visible to those parties not directly involved in the questions concerned.

From an overall point of view it is fair to state that this longstanding collaboration has contributed to place the Nordic countries among the most innovative, dynamic and highly developed telecommunication markets of the world.

Status and perspectives for the future

In 1998 the overall political and market situation of the telecommunication business is characterised by the following:

- A number of new services
- Increased liberalisation
- Political measures aiming at promoting competition (intention to have more players in the market)
- Changed regulatory schemes with the declared aim to have real competition as quick as possible, in some cases by introducing discriminating conditions for the incumbent (national) dominant operators
- Internationalisation – for business enterprises (our major customers) as well as for telecom operators
- Expanding market for traffic handling and for service provision
- Convergence between the IT market, the entertainment/cable TV industry and the conventional telecommunications market
- Technological convergence between the IT, telecommunications and the entertainment industry's technologies ('ICT technology')
- International alliances.

Taking this environment into account the Nordic telecom cooperation will therefore focus more on overall interests common to all professional players and less on matters related to operational questions which are of significance to the commercial and competitive situation of the individual enterprises.

Examples of such common areas of interests could be:

- Increased market volumes
- Harmonised business conditions (regulation)
- That customers do not perceive the range of products offered as non-coherent and unnecessarily complicated to use
- Synergy oriented co-operation in parallel with open mutual competition ('coopetition').

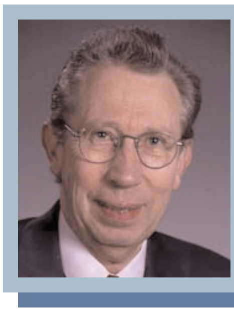
The main challenge for the new NORDTEL organisation in the years to come will be to further develop the activities within these areas of common interest and to increase the membership in order to create a solid basis for continued Nordic co-operation in the telecommunications field.

FOTO: FRIDOLFS BILDCENTER

Using the mobile phone in Nordic surroundings

Einar Utvik (58) is deputy director of Regulatory and International Affairs in Telenor AS. He is responsible for general and policy matters related to international organisations.

e-mail: einar.utvik@s.hk.telenor.no



Villy Toft (64) is deputy director of Tele Danmark's department for International Relations and Standardisation. Since 1992 his main area has been coordination of Tele Danmarks activities within ITU, ETSI, TRAC and other international consortia and fora. Villy Toft has since September 1996 been director of the Nordic Telecom Operators Business Association, NORDTEL (former "NORTEB").

e-mail: vtoft@tdk.dk