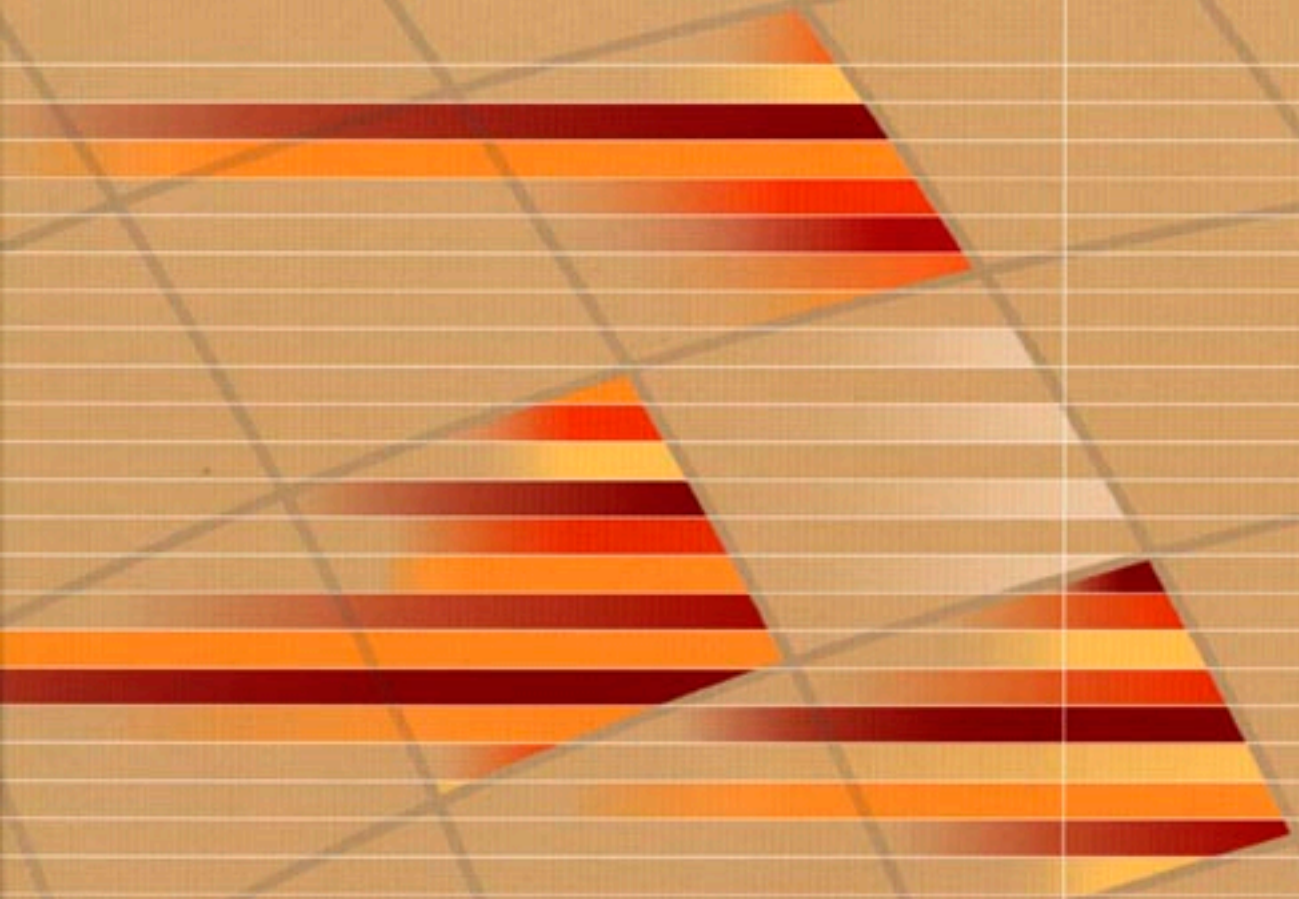
Telektronikk
2/2000

The Economy of Internet Services



Telektronikk

Volume 96 No. 2 – 2000
ISSN 0085-7130

Editor:
Ola Espvik
Tel: (+47) 63 84 88 83
e-mail: ola.espvik@telenor.com

Status section editor:
Per Hjalmar Lehne
Tel: (+47) 63 84 88 26
e-mail: per-hjalmar.lehne@telenor.com

Editorial assistant:
Gunhild Luke
Tel: (+47) 63 84 86 52
e-mail: gunhild.luke@telenor.com

Editorial office:
Telenor AS, Telenor R&D
PO Box 83
N-2027 Kjeller
Norway
Tel: (+47) 63 84 84 00
Fax: (+47) 63 81 00 76
e-mail: telektronikk@telenor.com

Graphic design:
Design Consult AS, Oslo

Layout and illustrations:
Gunhild Luke, Britt Kjus, Åse Aardal
(Telenor R&D)

Prepress:
ReclameService as, Oslo

Printing:
Optimal as, Oslo

Circulation:
4,000

Telenor

Contents

Feature:
The Economics of Internet Services

Special

Status

Kaleidoscope

Guest Editorial

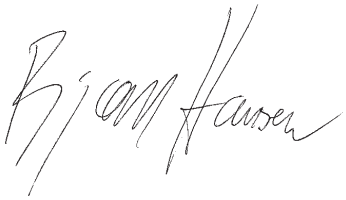
The Internet has over the last 10 years developed from a network connecting academic institutions to a network connecting households, the business sector and public institutions. Even people on the move can connect to the Internet via the cellular phone system. 10 years ago the ability to exchange E-mails was the main reason for connecting to the Internet. E-mail still exists, but the emergence of user friendly advanced World Wide Web applications has led to a vast number of new possibilities of exchanging information. This development has the potential of changing many aspects of society. Take business to business, E-commerce, as an example. Such a service will change business processes both within and between firms. The result is changes in transaction costs – eventually leading to a reorganising of firms and hopefully efficiency gains. In fact, the major part of the value contribution to society from the Internet may very well be due to such indirect effects. Nevertheless, such efficiency and welfare gains will only be realised if we are able to develop sustainable businesses of running a global ubiquitous Internet.

The ongoing development of the Internet is resulting in a whole series of business opportunities but also challenging questions for the telecommunications-, computer-, entertainment and information sectors in the economy. Some of these questions are purely technical, but most of them are of a hybrid technical-economic-strategic type, of which we will provide two examples in this feature edition of *Telektronikk*.

Our first example addresses quality differentiation of services – based on the proper technical solutions. In the strategic perspective of our example quality differentiation is a question of market segmentation, i.e. what combinations of price and quality will segment the market in the most profitable way.

Telecommunication firms have traditionally been vertically integrated. As new technology is introduced it becomes feasible to open the value chains. Why, where and how such an opening is implemented are clearly technical matters. However, an opening of the value chain will also have financial and business strategic effects. Therefore, our second example investigates the balance between vertical integration and bundling on the one hand, and the opening of value chains and unbundling on the other.

There could be numerous examples like the two mentioned above, where a successful overall strategy will have to combine technical knowledge of production possibilities with economic knowledge of the market. Traditionally engineers have studied optimal technical solutions for a given market structure, whereas economists have studied the market structure, segmentation and bundling, taking the production technology for given. The present feature section of *Telektronikk* contributes to bridging the gap between engineers and economists by presenting papers in which economic-strategic approaches to questions that hopefully are at the forefront of the technological development are addressed.



The Internet and the New Economy – an Introduction

BJØRN HANSEN

Bjørn Hansen (34) is Research Manager, Network Economics and Strategy at Telenor R&D. He is Cand.Oecon. from the University of Oslo (1992). Before joining Telenor he was a research associate at The Foundation for Research in Economics and Business administration at the University of Oslo. Current research interests are competition and business strategy in the converging Information, Computer and Telecom sectors.

bjorn.hansen@telenor.com

1 Background

The converging telecommunications, computer, information and entertainment industries (for short Information, Communication Technology, the ITC sector) are going through a period of dramatic change. Over the last five years or so we have observed a tremendous growth in the use of both the Internet and mobile phones. We have observed large increases in stock prices for dot com companies. At the third generation mobile (UMTS) licence auctions in 2000 the licenses were issued at unexpectedly high prices (630 and 615 Euro per capita in the UK and Germany respectively). We observe that players apparently give away services for free. The best known example is probably LINUX – a challenger to Microsoft Windows. Linux has been developed by voluntary contributions from the users. This operating system is available for free on the net. Furthermore, the technological progress is impressive. Finally, we observe that the market quite suddenly "explodes" in the sense that the usage of particular services increases very rapidly (e.g. SMS messaging in Norway).

As illustrated by the shares prices of dot com companies as well as the money raised in the UMTS auctions, both the capital market and central telecommunications firms expect that it will be very profitable to be players in the ICT markets. The strategic environment for these players in the very near future is however unknown. Apparently, many of the current business and revenue models are not sustainable. It is reasonable to expect this to change as new and innovative business and revenue models are created. At the current stage it is however difficult to assess the future profitability of a given service, and we cannot rule out the possibility that particular services have a significant market potential without being adopted due to limitations in the business and revenue models that are going to be established. In some cases we can have services that evidently have the potential of being adopted by many users, but where it is far from evident that it is possible (or optimal) to make money from it. Since we can rely upon experience from these markets only to a limited degree, economic theory becomes an important source of information.

Some are arguing that we are entering the era of the new economy and that the laws of this new economy will differ from the good old economy (see e.g. Kelly 1998). It is accordingly argued that one cannot use the traditional toolbox from economics when analysing business in the converging industry. In this paper we will however argue that the laws of the "old" economy still apply. The industry has characteristics similar to characteristics of other industries. The particular combination of characteristics may however be new, and market dynamics may accordingly differ from what we have observed in other markets.

In this paper we will give a brief overview of four[1)] fairly well known market characteristics that in combination can explain many phenomena in the so-called new economy. (See also Shapiro and Varian 1998 and Katz and Shapiro 1998.) The characteristics are:

- Rapid technological change;
- Economies of scale and scope;
- Strict complementarities;
- Network externalities.

The paper is accordingly organised as follows: In the following sections we will discuss each of the four characteristics. After the discussion we will give an overview of business strategy for providers of Internet connectivity by applying the characteristics since the Internet is very important for the development of the converging industry as well as being the focus area in this issue of *Telektronikk*.

2 Rapid Technological Change

Currently there is rapid technological progress in areas such as digitalisation, microelectronics, storage capacity, optical fibres, the capacity in routers, etc. The observation on advances in processor speed named "Moore's law" is an example of the rapid progress. Moore's law is basically the observation that the speed of processors (capacity) at the same price doubles every 18 months.[2)] This pattern has been fairly stable over the last 30 years. Another example is the transmission capacity in optical fibres. In the

---

[1)] *In addition to the four characteristics above one may also add* experience goods*, since the buying decision for many online services typically primarily will depend upon past experience with the particular service (see Shapiro and Varian 1998).*

period 1990 to 2000, new technology has resulted in more than a doubling of transmission capacity every year, even in already installed fibres. The technological progress is expected to result in reduced prices on current services as well as new services. This again is expected to result in new ways of economic interaction.

An important enabling factor in the convergence described in the previous section is that currently many technical devices are developed such that they are able to operate in an IP environment. IP is the protocol used in the current Internet, but it is also used in other networks. The IP protocol is a common "language" and thus all IP devices have the potential of communication over the same network – or over the interconnected ubiquitous networks that some argue will be developed over the coming years. This development can potentially change society quite radically. For example, many processes can be automated (the fridge can tell the grocer to send over more food, etc.). It is however far from evident that consumers adopt all such possible solutions. Nevertheless, there are potentially significant welfare gains in this development.

The technological development is on the one hand evidently resulting in increased welfare and is providing business opportunities, but on the other hand it is also resulting in a considerable uncertainty. This uncertainty is in particular about future production possibilities. Since the future evidently is unknown this uncertainty cannot be removed.

3 Economies of Scale and Scope

We can expect there to be economies of scale in different parts of the ICT sector. We can have economies of scale in the provision of information goods (content), in the provision of network connectivity, and in the production of terminals and equipment.

Communication networks are carrying digital signals. Varian and Shapiro (1998) define information goods as *anything that can be digitalized*. Information goods can accordingly be transported over the net. It is evidently significant economies of scale in providing such information goods. There are large costs attached to producing the first unit, but the cost of duplication or copying and distribution (over the net) is negligible.

Investments and the running of networks is typically characterised by both economies of scale and scope. As an example, a significant cost when building networks is civil works. This cost is independent of capacity being installed, thus there are economies of scale.

There can also be considerable economies of scale in producing some types of hardware. This can be traditional economies of scale or via learning by doing. Empirical studies indicate that the world market price on telecommunications equipment falls significantly with the total produced volume on the world market (see Olsen 1999).

The implication of economies of scale and scope is basically that size matters. In segments with economies of scale the biggest players will have the lowest average cost and thus have a comparative advantage. Furthermore, an optimal business strategy will typically be characterised by market segmentation. Such segmentation strategies are observed in other economies of scale industries like airlines and publishing. Finally, it must be taken into consideration that in "small" markets relative to the world market (e.g. Norway) it is particularly important to choose technology and technical solutions that also are adopted in other markets (standards as opposed to proprietary solutions) such that the small market (Norway) can take advantage of economies of scale in the production of such equipment.

4 Strict Complementarities

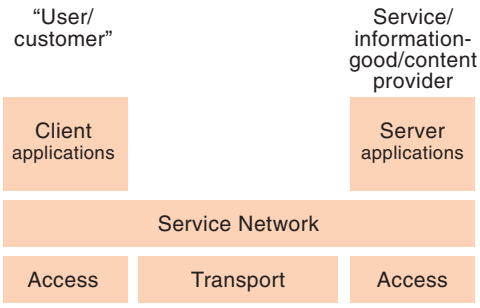
The demand from customers or "users" in the ICT industry is typically for composite services, i.e. services that are assembled from a series of separate services. An example is when a dial-up Internet user in Norway downloads some information from, say the US. This involves services from the local telephone company in Norway, the local Internet Service Provider (ISP), one or more Internet backbone providers, an ISP in the US and finally the information provider in the US. If one of the elements is removed, the information is unavailable for the user, and the other components in the chain become valueless to the user. Thus we have a structure with strict complementarities. In Figure 1 we provide a simplified illustration of complementarities.

In principle there are always at least two parties communicating over the network (human to human, human to machine etc.). In Figure 1 it is the client and the server that are communicating.

---

2) *This "law" is described in different ways. We can find examples at: http://www.zdwebopedia.com/Microprocessors/Moores_Law.html. All the descriptions of the law have the characteristics of fast and stable growth, and that is the feature that is important in our context.*

Figure 1 Complementarities



The two parties communicate by means of computers (terminals and/or servers) running software enabling them to communicate over a service network (e.g. the Internet). This network again is produced over a physical infrastructure consisting of e.g. access and transportation networks. The set of complements can be augmented in two directions. We can consider interconnection scenarios where there are several independent service networks that are interconnected. Furthermore, we can introduce more layers, e.g. on top of the Internet service layer we can have a bank running a payments network and above there we can have a retailer with a network of customers, etc. (On the layered industry, see e.g. Stabell and Fjellstad 1998.)

The presence of complementarities has impacts on business strategy. A classical example of business implications of complementarities is double marginalization (see Spengler 1950). Consider a composite of two services that is being supplied by two uncoordinated firms. Assume that both of the firms have market power. Then the firms will charge a mark-up on their component. It is straightforward to show that the total price of the composite good becomes (too) high and that the two firms can increase profits as well as increase the net value of the composite good for the consumers by a co-ordinated decrease of their prices. Furthermore, a provider of two or more elements in the composite good can price each element separately, he can price a bundle or he can restrict himself to setting a strictly positive price on one element and then give the rest away for free. Additionally, since consumer willingness to pay is for the composite and not for the single services, one single provider may extract all net profits from the composite good without sharing with the other providers of elements to the composite good. This is the case if one player has more bargaining power than the others, e.g. due to the control of essential facilities. Microsoft is an example of a firm that has deployed such strategies successfully. A discussion of impacts of complementarities in network industries is provided by Economides (1998).

5 Network Externalities

Utility from network participation depends on the number of potential communication partners and the quality of this communication. Accordingly, the willingness to pay for network participation is an increasing function of network size. Such externalities were first given a theoretical treatment by Rohlfs (1974). The strategic effect of network externalities on competition was recognized by Katz and Shapiro (1985), who analyze the strategies of choosing compatibility. As pointed out by Katz and Shapiro, externalities and the choice of compatibility are closely related.

Furthermore, complementarities discussed in Section 4 above may have effects very similar to network externalities. For example, it is possible that the variety of information services available over the network is increasing with the number of subscribers on the network. Then the willingness to pay for subscription will be an increasing function of the number of subscribers due to the effect via the complementary good, information services. This effect is by some called an indirect network externality.

Network externalities are of particular importance in new or growing markets. In such markets it may be necessary to sponsor usage, terminals or subscription in order to attract customers, and then after critical mass is being achieved increase prices such that the initial losses are covered. Furthermore, these externalities result in standards being of particular importance since the users' willingness to pay will increase if they expect that their interface/terminal etc. is compatible with services from other suppliers.

6 The Internet

The Internet used to be characterised by public initiatives, co-operation, peaceful coexistence and the users where characterised by a lack of commercial interests. There is reason to believe that the Internet would not have become what it is today without this pioneer period. Over the last five years or so, the Internet has been commercialised and there is an increasing conflict between private and public interests on the Net. The challenge is to keep sufficient openness, such that boundaries between different network domains do not hamper global connectivity. However, commercialisation is a prerequisite for further development of the network. The implementation of pricing signals on several layers will then be of vital importance.

There is reason to believe that the implementation of pricing signals will determine whether the Net will continue to develop as a digital tornado or whether it will end up as a second gener-

ation text TV, i.e. an electronic version of a postal service combined with an advanced bulletin board. It is possible to argue that the latter is the case. There are considerable challenges, or problems, with respect to technology and standardisation one has to solve before implementing pricing signals. Furthermore, there is reason to believe that many players on the current Internet would prefer a net that is less open and transparent. A commercial player with market power will typically consider it preferable to make boundaries between his domain and the rest of the Net visible to the user such that the player can differentiate its services from other services on the Net. When we combine the technical problems of implementing pricing signals with the commercial drive towards differentiation one possible scenario is that when new and advanced services such as broadband real-time applications (voice, video) only will be available over intranet like domains of the network. This is opposed to the open Internet where only "traditional" narrowband, non real-time services are available.

Even though the transmission capacity increases rapidly, we can expect that the combination of new real-time and broadband applications will lead to at least as rapid a growth in the demand for bandwidth.[3)] The Net is currently attracting new and impatient users (e.g. managers, brokers) and thus there is a need for more efficient ways to segment the market and allocate capacity. Otherwise, both users and applications being sensitive to delays will be excluded from the Internet. When capacity is a scarce resource, a negative externality is imposed on all the users in the same way as on a highway in periods of congestion. A user that is not facing a marginal price on his usage will not take into account the cost he imposes on all other users by imposing delays.

The price attached to Internet subscription and usage may serve two purposes. It is evidently a source of revenue for the service provider and furthermore, it may serve as a means to communicate to the user the true cost of using network resources. In the following we will focus on the latter, the price as a means to allocate scarce capacity.

Most applications running on the Internet use TCP/IP. This communication protocol is such that the network capacity is shared dynamically.[4)] The result is that in situations where capacity is scarce, the download time increases for all users. Furthermore, if one additional user starts to send packets, the download time is further increased. Thus we have a situation with negative externalities where each user is imposing waiting time on all other users. This is similar to the congestion experienced in busy periods on roads. As is well known from economics, if each user faces a marginal price on usage equal to the cost imposed on all other users (the opportunity cost), optimising behaviour by each user will result in the welfare maximising allocation of capacity (the first best solution).

6.1 Models of Internet Pricing

An Internet user is either paying a flat rate monthly fee for subscription or the price is a function of usage (time on-line / and or volume). Flat rate users are facing a marginal price of zero of initiating traffic on the network. These users are accordingly not taking into consideration the possible delay costs imposed on other users by starting to initiate traffic. Networks being priced according to this principle have to be dimensioned with large excess capacity or they can be expected to be over-congested in periods. Furthermore, there are no means of differentiation between different user groups implying that teenagers downloading entertainment material degrade service quality for business users. Flat rate pricing does accordingly have some drawbacks. As argued by Altmann et al. (2000) there is also empirical evidence indicating that users would prefer another pricing system such that quality is available for the users willing to pay for it. A major advantage of flat rate pricing is however that the required business support systems for network providers are inexpensive and very simple, as compared to the business support systems required for more sophisticated pricing. This is due to the fact that under flat rate pricing, the support system will only have to keep track of the number of subscribers.

The most common volume based pricing scheme is charging by the minute being on-line. For example in Norway, residential users of the Internet pay a price per minute of being connected to the ISP via the telephony access network. These users are accordingly facing a marginal price of Internet usage. By setting the price at the right level, the price can communicate the cost being

---

3) Taking PCs as an example, it seems like the increasing speed of the central processing unit and the growth of storage capacity is more than absorbed by the requirements from new software.

4) The applications not using TCP are typically using UDP, this protocol is more aggressive, it will not back off and reduce the transmission rate in the same way as TCP does when the network becomes congested.

imposed on all other network users and thus we may approach an optimal level of usage for the users being on-line. The marginal price is however attached to the time the user is being on-line and not the actual resource usage. The result of this pricing scheme is that users go off-line as soon as possible. This again can be expected to lead to lower network usage than what is optimal since it is time consuming and thus costly to log on and off the network.[5)] Time based pricing is by some seen as a major obstacle for the spread and usage of electronic services.

Under byte volume pricing the users are being charged by e.g. aggregated download volume or the sum of up- and download volume. There are examples of byte volume pricing in the Internet both in a commercial and an experimental setting. Under such pricing schemes, users can stay on-line without being charged anything unless they are active. Once they become active and start to use network resources, they are charged. Thus the drawback of time based pricing is avoided. In the same way as for time based pricing, by setting the price at the right level, the price can communicate the cost being imposed on all other network users. We can accordingly in principle implement an optimal solution by this pricing scheme. A possible drawback with byte volume pricing however is that users are "penalised" for the optimal behaviour of TCP in periods with excess capacity in the network. As described above, TCP takes advantage of periods with free capacity (periods where the opportunity cost is zero) by increasing the bit rate. Thus a TCP user will pay a higher price per unit time in uncongested periods as compared to congested periods. This is however not a drawback with byte volume pricing *per se*; it is a drawback attached to prices not varying dynamically. If prices are allowed to vary dynamically, the volume price will be zero in periods with free capacity and there is accordingly not a problem as described above.

Neither of the pricing schemes mentioned above discriminate between "local traffic" and "long distance" traffic. In the Internet long distance is typically traffic to and from the US. The required capacity in the trans-Atlantic links are very expensive as compared to local capacity and it may be of particular importance to introduce a marginal price on international traffic above the marginal price on local traffic. Such discrimination can e.g. be achieved by novel tariffing principles (see e.g. Oliwer and Songhurst 2000). Alternatively one can obtain differentiation by introducing a quality differentiation between local and international traffic, e.g. such that download time is lower for local traffic. Such a strategy is discussed by Foros and Hansen (2000).

7 Conclusions

In this paper we have argued that the combination of rapid technological change, economies of scale and scope, strict complementarities, as well as externalities are the main characteristics of the "new" economy. By adding (or emphasising) these features one can apply the standard toolbox of strategic analysis both for business and regulatory purposes.

Literature

Altmann, J, Rupp, B, Varaiya, P. 2000. Quality Matters : Some Remarks on Internet Service Provisioning and Tariff Design. *Telektronikk*, 96 (3), 20–25. (This issue.)

Choi, S Y, Stahl, D O, Whinston, A B. 1997. *The economics of Electronic Commerce*. Indianapolis, Ind., Macmillan Technical publ.

Economides, N. 1998. *Raising Rivals' Costs in Complementary Goods Markets : LECs Entering into Long Distance and Microsoft Bundling Internet Explorer*. Stern School of Business, N.Y.U. (Discussion Paper EC 98-03.) Available at http://www.stern.nyu.edu/networks/98-03.pdf

Foros, Ø, Hansen, B. 2000. Interconnection and Competition Between Portals Offering Broadband Access. *Telektronikk*, 96 (3), 26–37. (This issue.)

Foros, Ø. 1998. *Internett : Digital tornado eller 2. generasjons tekst-tv*. Kjeller, Telenor R&D. (Report R&D R 26/98.)

[5)] *Notice however that since users log off the network as soon as possible, the required number of IP addresses is lower under time based pricing as compared to e.g. flat rate pricing since the network typically only allocates addresses to the users being on-line. The address space of the currently used IPv4 is limited and is by some expected to be a scarce resource. When/if the number of addresses becomes a limiting factor, time based pricing may turn out to be optimal. Alternatively, one can upgrade to IPv6 which has a much larger address space. This upgrade will lead to non-trivial problems of co-ordination.*

Katz, M, Shapiro, C. 1985. Network Externalities, Competition and Compatibility. *American Economic Review,* 75, 424–440.

Antitrust in software markets. (2000, September 13) [online] – URL: http://www.haas.berkeley.edu/~shapiro/software.pdf

Kelly, K. 1998. *New Rules for the New Economy : 10 Radical Strategies for a Connected World*. New York, Penguin.

Klemperer, P. 2000. *What really matters in auction design*. Paper presented at the 11th European regional ITS conference, Lausanne, 9–11 September.

Oliver, H, Songhurst, D. 2000. Market Managed Multiservice Internet. *Telektronikk*, 96 (3), 38–44. (This issue.)

Olsen, B T. 1999. OPTIMUM – a techno-economic tool. *Telektronikk*, 95 (2/3), 239–250.

Rohlfs, J. 1974. A Theory of Independent Demand for Communications Service. *Bell Journal of Economics,* 5, 16–37.

Shapiro, C, Varian, H. 1998. *Information Rules : A Strategic Guide to the Network Economy*. Boston, Mass., Harvard Business School Press.

Spengler. 1950. Vertical Integration and Antitrust Policy. *Journal of Political Economy,* 58, 347–352.

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

A Business Model for Electronic Commerce

LEIF B. METHLIE

Leif B. Methlie (61) is Professor of Information Management at Department of Strategy and Management, Norwegian School of Economics and Business Administration (NHH), Bergen. Methlie has taught and conducted research on information systems to support management and decision makers based on theoretical frameworks within management, cognitive sciences and systems theory. From 1996 electronic commerce has been his major research area and he is currently managing several research projects concerning new business models in the network economy. More information on these projects can be found on http://emarkets.nhh.no. Methlie teaches graduate courses on electronic commerce, decision support tools, and knowledge management. Methlie is a former President of NHH.

leif.methlie@nhh.no

Introduction

Electronic commerce has existed for several decades with interorganizational information systems, electronic document interchange (EDI) and electronic payment systems. However, these systems have been closed proprietary systems where mostly large companies could afford to participate. The real break-through of electronic commerce came with Internet and the World Wide Web. By 1993 Internet had reached a critical mass, a stage where the power of attraction increased tremendously. Internet became the platform of commercial activity and new electronic markets emerged. The Internet technology established a platform accessible through telecommunication networks all over the globe. Information systems were no longer constrained to well-defined databases accessible only through proprietary networks. The real open-ended information system was born where the user could navigate on the web by context-defined hypertext links; thus no predefined physical network links are required.

Business has discovered the enormous power of the web. As electronic commerce grows on the web more business will enter the network. Very few, however, seem to be prepared for a successful transition to the web. For many companies the web is regarded as just another delivery channel, offering customers an additional channel of services, but lacking a strategy to pursue new business opportunities. Business in the marketplace is copied into the virtual world, the marketspace. In this paper we shall see that being competitive on the web requires new business models and strategies. The main business functions like marketing, sales and distribution are the same, but the logic behind them is changed. Internet is a *mediation technology* that enables new distribution channels, new marketing communication opportunities and new trade environments to be created. Market transactions are exchanged in an open electronic information network based on standardized communication protocols. Furthermore, Internet is a two-way interactive network enabling millions of people to engage in dialogues. Information or transactions may be pushed by the supply side or pulled by the demand side. Internet offers a channel of high multimedia capacity and long reach, a feature rather rare in marketing where trade-offs normally have to be made between reach and richness (Evans and Wurster, 1997). These properties of the Internet make it more than a technology – it is a new business concept.

Although the penetration of electronic commerce is still low we see a new economic landscape emerge based on new, governing principles – the network economy. This economy will have profound effects on both the demand side and the supply side of the markets. New business models have to be developed taking into account the fundamental shift in economics of information that follows from networks such as the Internet, and the explosion in connectivity that leads to new customer values, disintegration of value chains, and disintermediation of distribution systems. These changes are taking place vertically in the value chain. However, the new medium also creates opportunities for horizontal aggregation of resources (content and customers). How do we arrive at the new business models? What are the new value propositions needed to meet customers buying values; is there an evolutionary track to follow; and what is the strategic platform? These are some of the questions addressed in this paper. The objective is to present a *business model* and some *business strategies* that follow from this model in order to stay competitive in this new business environment.

The discussion will start by looking at the impacts of the new economic environment on businesses. After this follows a section on new customer values in the marketplace. A business model is then described resulting in a strategic framework for Internet implementation of electronic commerce. Figure 1 illustrates the content of the paper.

Value Creation in Networks – the Network Economy

The driving force behind the digital revolution is the remarkable science of microprocessors, which has shifted the market economy from an industrial-based to an information-based economy. The exponential improvements in microprocessor speed, size, and cost that have operated since the 1960s follow Moore's Law: Every eighteen months, capacity of the microprocessor doubles while production costs are kept constant. The result of this dramatic fall in price/performance is the ubiquity of the microprocessor.

Less known than Moore's Law is the observations made by Robert Metcalfe that networks dramatically increase in value with each additional node or member. He formulated these observations in a heuristic model, also known as a law: The usefulness, or utility, of a network equals the square of the number of users. The more people who use your software, your network, your standard – the more valuable it becomes; and the more new users it attracts, the more its utility increases and also the speed of its adoption by still more users (Downes and Mui, 1998).

Moore's Law made possible the digitization of nearly everything surrounding us: toys, appliances, telephones, credit cards and cars. Metcalfe's Law puts focus on network values. Lack of standardized communication protocols, however, has made the connection of all these devices a problem. The break-through came with the Internet and the communication protocol TCP/IP. A global network of millions of connected computers was created. With decreasing communication costs we can foresee vast networks of connected computers and non-computer chips. *Global connectivity* is a reality. This explosion in connectivity will lead to new economics of information changing the structures of entire industries and the way companies compete. In this new economy value creation occurs in networks and values are created in connections. Metcalfe observed the value of networks – economists denote this the *network externality*. A positive externality is the benefit or value of an asset not due to its value as such, but to the value of adding the asset to the network, i.e. adding one more asset to the network directly effects the value of all other assets connected. All communication technologies exhibit externality effects. A telephone alone is not of much value. The real value of a telephone is its ability to establish connections for communication and not the cost of the equipment itself. The more telephones connected, the higher is the value of each phone. Thus connections create values. Since one new telephone added to the network connects to all existing telephones in the network, value creation is *exponential*. This exponential growth is a significant property of the network economy: value grows exponentially with membership, while this value explosion sucks in yet more members. Exponential value creation favors the early entrants and may lead to monopolies.

Success companies like Microsoft and FedEx and success technologies like fax machines and the Internet, all show this exponential growth property, albeit a very slow start. During the first ten years, Microsoft's profits were negligible. The trajectory of FedEx is similar. The fax machine that was invented in 1843 and introduced as a service in the United States in 1925, remained a niche product until 1980. Then, in the course of five years, the demand exploded. In the network economy, values created by memberships can be obtained by any business on the web as will be shown below. Related to the exponential growth curve of network companies is the tipping point, the point after which success feeds itself (Kelly, 1997). In the network economy the tipping point is lowered, and what is more important, this point is reached before the market realizes its significance. Microsoft and FedEx are good examples. The lesson to be learned is that success in the network economy is nonlinear and it favors the early movers. To wait and see may be too late. On the other hand, a slow start may be more capital demanding than most of the entrants seem to realize.

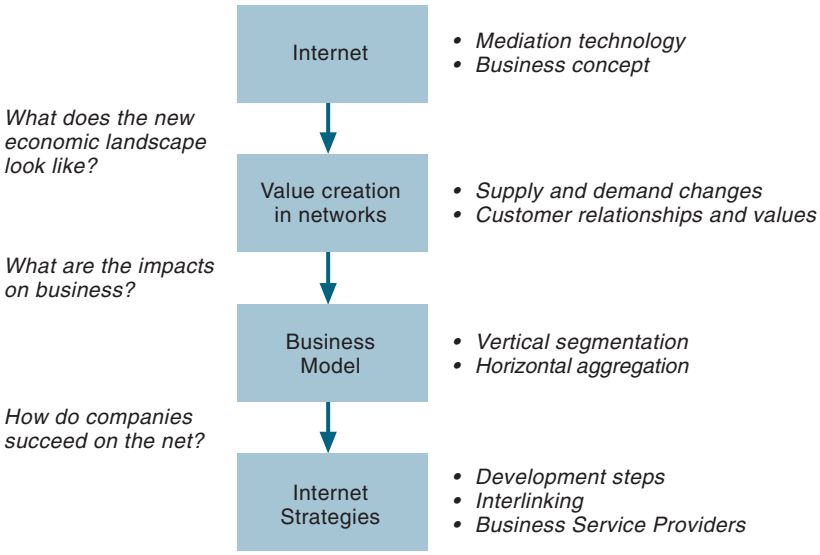


Figure 1 A presentation of the content

There is a corollary to this. If the network value increases with the members of the network, and this value increases exponentially, the price of these entities creating positive network externalities should be low, or even free. Income should be earned on related products and services. This is exactly what can be observed. Mobile telephones are almost given away to increase the population of subscribers. Money is earned on its use. Internet browsers are given away while associated services are charged. The result is a supply driven demand growth.

The industrial economy of scale stems from the effort of a single organization producing more for less. The values created belong to a single company. In the network economy, increasing returns are created and shared by the entire network. According to Hagel et al. (1996), a company's primary focus in the network economy shifts from maximizing the firm's value to maxi-

mizing the value of the infrastructure (the network). Porter's theory of competitive positioning (1980,1985) put network values on clusters. On the web, however, relations are tightly coupled but loosely formalized. The web-clusters are therefore more informal. Standards are the means of formalizing and strengthening the relations on the networks, and the companies at the gateway to a standard are likely to win. The tightly coupled network creates a turbulent passover from one success to another. Long term stability requires continuous renewal. Economy of speed becomes an important factor in the new economy.

Information technology seen as a production technology has been thought of as bringing supreme productivity. In recent years some economists have challenged this statement. But whether the technology has created improved productivity or not, it has indubitably led to new things, new services and new values. In the new economy we should worry less about doing all the old things right, that is, to be more productive. Rather the important issue is to do the right things. On the web the bottleneck is not productivity; it is creativity and imagination. Being innovative focusing on what adds value to the customers is what counts. George Gilder puts it this way, «Don't solve problems, seek opportunities» (Kelly, 1997).

From the characteristics of the new economic environment of networks some general conclusions on both the demand side and the supply side can be drawn. On the demand side we see new customer values emerging with access to electronic networks. Competition will increase and prices will fall, customers' demand is stimulated by free offerings, and the media becomes the market. In fact, all commerce will be performed electronically on the web, that is, sales and buying will be executed on the net; distribution of physical things will, of course, have to be done outside the net. On the supply side the changes will occur in the value creation processes. These processes must, however, match with the new customer values created by the networks, thus leading to new customer relationships, new business models, and new economies of scale, scope and speed. In the new environment, competition will be more innovation driven.

Customer Relationships and Customer Values in Marketspace

Value Creation Components

In an article in Harvard Business Review, Rayport and Sviokla (1994) introduced the term «marketspace» to designate an environment in which buying transactions are executed on electronic networks. Marketspace is the virtual substitute of the marketplace where seller and buyer meet physically to execute a transaction. In marketspace physical (real world) entities are substituted by their symbolic representations, that is, by information about these entities. Thus a buying transaction is based on informational entities (data, text, voice, images and video). For some entities we trade with their informational representations only, e.g. paying bills, buying and selling stocks, entertainment and news. In these cases, not only buying transactions are settled in the marketspace, but delivery can also take place here. In the case of goods, i.e. physical products, e.g. groceries, cars or computers, delivery will take place at the marketplace.

Today you can buy a PC on the Internet. Most of the vendors present their products well specified on all relevant attributes. You can fill the order form and even make or organize for the payment on the same screen. This is an example where the PC itself is substituted by information about the PC and where the buying transaction takes place on the network, i.e. the marketspace. However, there is more to a product than the physical thing itself, such as information and services. Presale activities include marketing information and negotiations, sales include delivery and payment, and after-sale activities include training, maintenance, and warranties. To understand how customer values are created in the marketspace it is important to recognize the various components of a product that create these values.

Rayport and Sviokla claim that values are not only created by the product and services offered by the vendor, here called *content*, but also on the *infrastructure* that enables the transaction to occur, and the *context*: the way the vendor is offering its products and services. Furthermore, they claim that it is important to disaggregate the conventional marketplace value proposition «to create new ways of adding value, lowering costs, forging relationships with 'nontraditional' partners, and rethinking, 'ownership' issues ... Information technology adds or alters content, changes the context of interaction, and enables the delivery of varied content and a variety of contexts over different infrastructures» (Rayport and Sviokla, 1994, p.145).

Since value creation is not only associated with the content itself we have to examine the infrastructure that facilitates transaction exchange in the market and the context in which customers interact with the vendor.

Supplier – Customer Interaction

The interactions between vendors and customers are determined by properties of content, infrastructure and context. An electronic network like the Internet is an *interactive* medium where vendors and customers can communicate online; where vendors can advertise their offerings and customers can advertise their demands (reverse markets); and make transactions. It is a many-to-many communication network where customers can search among offerings from multiple vendors, and vendors can reach many customers by an information rich channel (multimedia). Interactive channels enable customers to easily provide feed-back to vendors. Also suppliers can provide interactive support facilities to customers (help facilities, education and training, etc.). Furthermore, customizing of products and services to particular customer needs is possible; likewise, increased customer participation (self service) in the buying processes.

Another feature of the Internet as a transaction medium is customer *transparency* of solutions (content). The concept of transparency hinges on the capability of the web to interlink web-sites and provide customers with complete solutions where the needed content is provided by several suppliers, e.g. travel or real estate buying. Network linking capitalizes on both economy of scale and scope and is one of the success factors of businesses performing electronic commerce.

Many-to-many communication enables another new feature of networking, the establishment of *virtual communities*. Virtual communities enable people with a common interest to interconnect and exchange ideas and experience. «The rise of virtual communities in on-line networks has set in motion an unprecedented shift in power from vendors of goods and services to the customers who buy them. Vendors who understand this transfer of power and choose to capitalize on it by organizing virtual communities will be richly rewarded with both peerless customer loyalty and impressive economic return.» (Hagel and Armstrong, 1997, p 2). Virtual communities set up other channels of communication. Here people are aggregated horizontally, and customers, for example, can exchange ideas and experiences among each other, or as a group they may communicate with the vendor (brand community).

Another feature of the technology is the capability of *push and pull* communication. Push-technology enables the vendor to push information to the customer whenever the value of the desired information is changed (stock portfolio status, price changes, etc.). On the other hand, customers can pull information from vendors on demand.

The context of electronic networks provides the customers with a new *interface*. Access and information provision change. Multimedia enables information representations of various types to be presented to the user. Information on demand allows content disaggregated and bundled in different ways.

New Customer Values

Ultimately, buyers have the power to determine the business success of a supplier. Therefore, it is of utmost importance to understand the values created for the customers by the products and services provided. These customer values are known as the supplier's value proposition. It refers to the business opportunity between a business and a customer (Gascoyne and Ozcubukcu, 1997). When a business moves into the net for electronic commerce new customer values are created that change the value proposition of that business. Respecting the value proposition requires placing the customer at the center when deciding on the business strategies. A critical question to ask is how well does the current value proposition serve customers in the marketspace.

Central to building a competitive value proposition for a business in the marketspace is to understand the values that can be created by this medium. By analyzing the value creation components: content, infrastructure and context, along with the new customer relationships that emerge we have arrived at the following values:

Content-related values (intangibles):
- customization/personalization of services
- proactive, anticipative services
- reverse markets or advertising
- learning programs
- experience sharing among customers

Infrastructure-related values:
- on-line
- speed
- interactivity
- connectivity
- convenience
- reduced search costs
- trust

Context-related values:
- decision tools
- multimedia

Intangibles – Content-related Values

Increased personalization of products. The network enables the supplier to customize products and services to its customers because more information can be collected about each individual customer. This enables customer profiles to be established and products to be personalized. Another important customer value is the Internet capability of pushing information from suppliers to customers. By the push technology customers receive updates or alerts as conditions predefined by that customer occur. Typical push information is stock prices, trading volumes, etc.

Proactive, anticipative services. Furthermore, the network provides the opportunity for a supplier being proactive in assisting customers through their various stages of product acquisition and utilization; from requirements evaluation, through acquisition and stewardship, to retirement (cf. the customer resource life cycle described in Learmonth and Ives, 1987). The supplier also has more information about customers and can therefore be more anticipative of customer needs and desires.

Proactive buying – 'reverse advertising'. Bulletin boards or auctioning software will allow customers to announce their product requirements and accept bids. Intermediaries such as Priceline.com will connect buyers to sellers on the customers initiative. Interactive communication can also be utilized to 'include' customers in business developments, for instance by participative product design.

Learning facilities. Associated with building relationships with customers by providing more information on products and their usage, one step further is to provide learning modules with the products that put the customer in a better position to understand and utilize the product or service provided.

Experience exchange – learning from other customers. Internet allows various kinds of interactions among users. It provides discussion forums where users can share information and create 'knowledge bases'. One popular knowledge base is 'frequently asked questions' or 'FAQs'. Discussion forums can be public or private. Customers may, for instance, form groups around brands, so-called brand communities. This enables them to share experiences and communicate with the supplier as a group. It can be organized publicly around themes or privately by special interest groups. Virtual communities are treated further in the section below on horizontal aggregation of customers.

Infrastructure-related Values

On-line. The network enables the customers to be on-line with its suppliers. Furthermore, the communication channel set up is **interactive** offering the opportunity of dialogs on the net.

Connectivity. With no physical barriers the customer can scan the world market on the net. For information products the cost of supply is almost independent of distance. Even for goods ordered on the net, new delivery services have reduced the barriers of distance. For example, an automobile owner in Iceland used DealerNet to find a dealer in Seattle who could supply a part for his Nissan Pathfinder (Gascoyne and Ozcubukcu, 1997, p. 38).

Convenience – anytime, anywhere access. The net is open 24 hours a day, every day through the whole year. Access can be done from any client connected to the Internet anywhere at any time. Thus one has greater freedom in choosing a convenient time and place for shopping, paying bills, etc. than in the traditional world. Furthermore, there are no physical distances involved in moving from one outlet to another on the net. New access providers such as America Online and Scandinavia Online will further increase the convenience of shoppers by setting up storefronts that host content providers.

Reduced search costs – more information. Searching the net for suppliers of needed products and services is fast and cheap. Economic theory tells us that reduced search costs lead to search for more information on alternatives and their properties. For standard products competition will increase leading to reduced prices and reduced price variations in the market. For differentiated products the quality will increase. More information will be collected before a decision is made, and a better-informed customer is supposed to make better decisions. Buying behavior is expected to change towards more emphasis on specific product attributes rather than brand. Going back to our example of buying a PC on the net, product information can be found on the suppliers web sites and easily compared. However, if intangible attributes like quality, service, reputation, etc. are important factors for the buying decision, brand name may have a greater effect on the selection process than product attributes. So far we have little research on changing buying behavior due to lower search costs and more informed customers.

Trust. Market transactions require trust among the participants in order to avoid opportunistic behavior. Customers encounter greater risks

buying in electronic markets due to legal aspects such as warranties and product failure when buying across different jurisdictions, payment security, authentication, and privacy. To provide customers with trust is recognized to be of very high value in electronic networks to prevent suppliers not fulfilling their obligations in market transactions.

Context-related Values

More evaluation – better decision tools. Software either located at the supplier's server or in the client can assist the customer in deciding on alternatives. A typical example is a calculation program for pay back and interests of a loan. This program can either be provided by the lender or it may be part of a personal financial system on the customer's desktop. In the latter case they may be able to draw on data from multiple institutions and compare alternative product offerings. The sheer breadth of choice available on the net will create the need for intermediaries to play the role of facilitating agents. Agents are software programs that search on the web and evaluate products and services. They may operate as independent intermediaries and establish databases that guide users in specific market or product segments (e.g. loan guides, wine guides, travel guides, etc.). Agents may also be resident on or are initiated from the client.

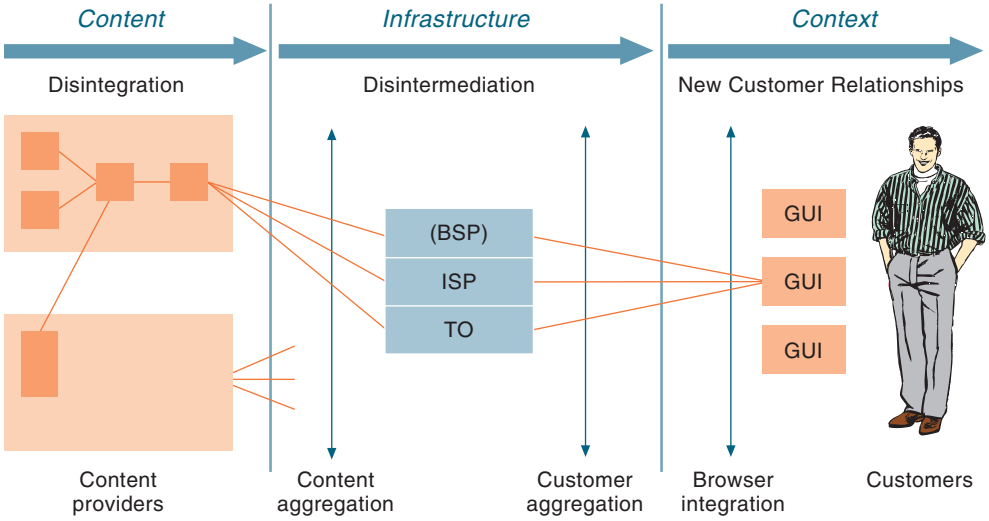
Multimedia. The ability of the Internet to treat various information representations as digital signals provides suppliers with richer channels and a variety of presentation forms.

These new customer values must be built into the customer value proposition of the vendor in order to be competitive on the web. The main conclusion to be drawn so far is that the *power of the consumer* will increase with business on the net.

A Business Model

The basis for our business model for electronic commerce is the vertical segmentation of the value chain into content; infrastructure and context as proposed by Rayport and Sviokla (1994). We shall discuss vertical segmentation of the value chain further below.

Telemediated commerce, however, requires channels of communication for delivering content from the supplier to the customer. These channels become a new market for telecom operators and value-added service providers – a market which is different from the traditional telecommunication market of setting up connections between users. The new market is a consumer market where the distinction between activities such as creation, processing, distribution and presentation of content is disappearing. This new, converging market is extremely complex with the content providers on the one side, the consumers on the other, and with a complex, multi-layered value network in between. We shall differentiate between two types of value-added service providers: 1) content aggregators, who we shall call business service providers; and 2) value-added telecom service providers, who in our context normally are called Internet service providers. Finally, there are the network operators – the transport providers such as telephone, cable and television companies. It should be noted that the business service provider is an optional intermediary. Content providers may deliver their content directly to the consumer market through the channel and the end-user's



Legend:

BSP: Business Service Provider
ISP: Internet Service Provider
TO: Tele-operator
GUI: Graphical User Interface

Figure 2 Components of a business model

equipment. The three-layered distribution channel chosen here is arbitrary. In a fully deregulated telecommunication market the telecommunication channel may be split horizontally into several more layers and vertically upstream or downstream, on the basis of roles played in the mediation process. How many separate players or businesses that will eventually emerge on the arena is difficult to foresee at this stage. The organization and operation of these networks are, among other things, contingent on standardization of the interfaces.

Electronic networks provide global connections. This property of networks can be used to establish new communication patterns among users, both suppliers and customers that can be exploited for business values. Linking of suppliers can aggregate content; suppliers can communicate interactively with their customers, and customers can communicate with one another and form communities. Aggregation of content and of customers will be dealt with below.

A business model for electronic commerce must have all these three features as described above. The main components of our new model are depicted in Figure 2.

Vertical Segmentation – Decomposition of the Value Chain

The new medium of electronic sales and distribution channels has different characteristics than traditional channels. Not only do these new channels offer fast, direct and low cost communication. They also offer unique capabilities in terms of new customer values as described above. To look at this as just another cheap channel is to underestimate the potential and the threats because the network will precipitate changes in the structure of entire industries and in the way companies compete. Banks, for instance, have traditionally owned (branches) or at least controlled (telephone or proprietary online, terminal systems) their distribution channels. For these channels they can decide on the user-interface themselves. They are the number one contact point for their customers. In the marketspace banks are facing a new situation: they no longer are the number one contact point. The bank may be the number five! (Cf. the business model in Figure 2.) This picture is a radical, though readily feasible, departure from the traditional view of a bank providing services and products to its customers with which it has a dedicated link without the intermediation of other participants in the value chain. Also the traditional value chain for physical goods undergoes changes. Virtual display of products does not require physical closeness to customers. Thus the functions of wholesalers and retailers will change. The product can be virtually displayed and sold by the producer, thus savings in distribution costs may be high (see for instance Benjamin and Wigand (1995)). In the physical value chain intermediaries may disappear. The structural changes have therefore been termed *disintermediation*. However, as shown in the bank example, new players are entering into the distribution channel. Therefore, the best term for the changes taking place may be *reintermediation*.

Vertical segmentation implies that the value chain, from the content provider to its customers, is decomposed and new intermediaries are emerging. At the customer side we find the *context provider* who provides the user interface to the network. Software firms like Microsoft with Windows, browser and special purpose software like Money for personal finance and bank transactions, are positioning themselves with strong branding, large customer bases and expertise. Generalized (standardized) software offers the customers more variety in their buying choices than proprietary solutions. The advantage of the context providers is their close contact with a large customer base. They are the number one contact for the customers when they log on to the network

So far we have looked at the vertical segmentation of the distribution channel – disintermediation and reintermediation. However, vertical segmentation also applies to production: disintegration and reintegration. By disintegration we mean breaking up an integrated production process and outsourcing parts of it taking advantage of lower transaction costs and increased scale economies due to the information technology. Disintegration of production also lowers the barriers of entry for special process providers allowing new entrants into established industries. Disintegration requires reintegration for providers of customer solutions. New entrants with effective management of customer relationships may outperform incumbents by providing solutions assembled from a number of cost effective suppliers or subcontractors. Banks may function as an illustrative example again. Production of bank products and services has traditionally been an integrated, internal value chain. A product such as a loan is produced through a series of subprocesses: sales, information collection, credit evaluation, management of collateral, etc. Each of these processes has its own scale of economy and may be outsourced. Contract banking, where a seller of banking services contracts with several suppliers of banking processes, is one direction banks may be moving with the new technology in place and fully utilized. However, external pressure from new entrants to the banking arena seems to be necessary for this change to start.

The newspaper trade is another industry where it is easy to see a disintegration of the value chain. Journalists may communicate directly with their readers on different topics. Pieces of the newspaper can be unbundled. What happens to the newspaper companies' economy if classified ads, which account for 25 % of a typical newspaper's revenues but less than 10 % of its costs, are separated from the rest? Bundling of services is often a cost accounting problem leading to cross-subsidizing and bad pricing strategies; thus making incumbents vulnerable for cost effective, new entrants entering into the electronic environment exploiting the unique capabilities of electronic networks. Disintegration raises several new issues. Pursuing core competencies will be more important and web strategies linking businesses will be critical for success. That is where value is created and competitiveness is sustained. For traditional companies moving into the network a major issue is what kind of role it wants to play on the web. In a disintegrated environment it may take the role of an integrator – a business service provider or a web host aggregating content. However, this requires a very offensive early mover strategy which incumbents hesitate to take unless competition from the outside forces them to do so.

Multi-layered Networks

As described above, telemediated sales and distribution introduce new players into the value chain. Our framework, as shown in Figure 2, consists of a three-layered distribution network, although the top layer, the business service provider, is optional. The choice of a three-layered network is based on what we observe as major players in electronic commerce today. Several roles of intermediaries can be identified. Roles are, however, not equivalent to companies or businesses. In this section we shall primarily be concerned with the business service providers.

Delivering services in the marketspace is just as much bundling of offers as single products. Services may be provided by one single company, or composed of content produced by several, specialized companies. Companies that aggregate content (bundling) for customers is called *business service providers (BSPs)*, also known as integrators or portals. The primary function of a BSP is to bundle content for convenience, price and quality (see below). A BSP has the same role in the marketspace as the retailer or a broker in the physical marketplace. It serves as an intermediary between customers and suppliers. Resnick et al. (1995) suggest that brokers are important in a market because search costs, lack of privacy, incomplete information, contracting risk, and pricing are better managed through a broker. Malone et al. define similar reasons to have intermediaries: reducing coordination costs, addressing problems of asset specificity, and promoting standardization. Bailey and Bakos (1997) define four roles of market intermediaries:

a) *aggregate* buyer demand or seller products to achieve economies of scale and scope and to reduce bargaining asymmetry;

b) protect buyers and sellers from the opportunistic behavior of other participants in a market by becoming an agent of *trust*;

c) *facilitate* the market by reducing operating costs; and,

d) *match* buyers and sellers.

In addition to the roles identified by Bailey and Bakos we may add managing customer relationships, serving as a clearinghouse for payments, and providing a single format for product presentation and pricing.

Major players as business service providers in the US are, among others, America Online, Compuserve and MSN (Microsoft Network). In the UK we find Virgin Online and in Scandinavia Scandinavia Online (SOL). These players have strongly branded names. MSN have exclusive branded channels such as Microsoft Expedia, a travel and online booking service in USA and Canada with links to 10,000 selected sites (Dedman, 1997).

To support the rapidly growing commercial Internet, an entirely new industry called Internet service providers has emerged. The Internet service provider industry offers access to the Internet (domain names), connectivity and routing on the basis of Internet protocols (TCP/IP), provides client and server software for navigating and publishing content on the Internet, and network management including accounting procedures for charges.

At the most technical and content remote layer we find the telecom operators who provide the infrastructure and transportation of signals. They provide the subscriber access network, that is connectivity to end user equipment.

Horizontal Aggregation

A distinctive value of interactive electronic networks is the ability to aggregate resources and form communities. These properties of the network have been described earlier (see for instance Winer (1996), Hagel et al. (1996) and Hagel and Armstrong (1997)).

According to Hagel et al. (1996) user values can be derived from three kinds of aggregation: 1) aggregation of users; 2) bundling for convenience and cost; and 3) bundling for quality. The two latter types are aspects of content while the former in a business context will be related to customers.

Aggregation of customers. Electronic networks provide the unique ability to interconnect users to form communities which can have their own voices. The Internet has several facilities for user interactions: e-mail, news groups and chat rooms (Internet Relay Chat). News groups are organized by bulletin boards and anyone can set up a newsgroup by providing a bulletin board for exchanging views on a particular topic. This part of the Internet is very fragmented and its size is difficult to estimate, but according to Hagel et al. (1996) there are as many as 60,000 bulletin boards serving some 6 to 10 million members. Similarly, the most popular service provided by the Internet is e-mail, and also the Internet Relay Chat service has acquired many users. Common to these services is the creation of groups or communities by interested subject.

If we put these services into a business concept, we now have a powerful facility for exchange of experiences and ideas around content (products), brands, interests or needs. Who should organize these communities? The content provider to strengthen the customer relationships can do it. Independent service providers growing out of specific consumer interests may also do it.

One brand community is that of Saab owners. Owners of the Saab car are notorious for their fierce loyalty to the brand and their high degree of interest in new product offerings and other matters related to the company. Saab owners have their own 'chat rooms' where they discuss the cars with each other and communicate their feelings to the company via e-mail. Winner et al. (1996) raise several interesting research issues emerging from a marketing point of view stemming from organizing users in this kind of collective. Among the issues are how membership affects purchasing behavior; around which products are communities likely to form; and can a community be regarded and treated as a market segment?

A content provider, a business service provider or an association of consumer interests can organize a community. For a content provider, the goal could be to strengthen its brand image and establish closer ties with its customers. It may cooperate with the community members on issues around product design and development, presenting early news and get feedback on customers' experiences with the products. A community may have both positive and negative effects on its members. Positive feedback to the producer strengthens its brand. Negative feedback announced for the whole community may create a collective negative attitude, amplified compared to individual experiences. However, business can only succeed by satisfied customers!

Frequently asked questions (FAQs) are often the starting point for aggregating customers' views. Collecting questions received by e-mail from customers, organize them and make them available to anyone visiting the web-site, is a way of aggregating customers' views. It provides, however, no direct communication among the customers. One problem with FAQs organized by a supplier is that of biased selection of questions and of supplier constructed questions.

Direct communication can be accomplished using bulletin boards where customers can post messages, comment on topics, etc. Chat channels are one of the most popular activities on community bulletin boards.

Aggregation by content. Bundling of content has the objective of providing customers increased convenience and quality, and low prices similar to what we find in the retail concept in the physical marketplace. Content falls naturally into clusters like groceries, travel, financial services, etc. «For most users, having a full range of resources bundled together in one accessible format with a consistent look and feel and integrated billing will be much more convenient than surfing the Net to assemble an equivalent set of resources and wrestling with format incompatibilities and disparate billing systems.» (Hagel et al., 1996). Retailing on the net is performed by web integrators or business service providers (BSPs) as described above.

Value networks. Stabell and Fjeldstad (1998) have defined three generic value configurations based on Thompson's (1967) typology of long-linked, intensive and mediating technologies. A value network models firms that create value by facilitating a network relationship among their customers using mediating technology. A business service provider is a mediator linking content providers by aggregating content or linking customers into communities. The primary activities of a value network are network promotion and contract management, service provisioning, and network infrastructure operation. Value network services are characterized by demand side economies of scale resulting from positive network externalities (Katz and Shapiro, 1985).

Internet is an enormous resource base. But how do we find the best offerings we are looking for? Search engines and agents are not well enough developed at the moment. The retail concept may be the best solution to find what you are looking for. Retailing may also be price favorable and quality optimal since a BSP may be more knowledgeable about offerings on the net and may also be able to negotiate with suppliers.

Internet Strategies for Electronic Commerce

Above, we have discussed how the network economy creates changes in both supply and demand characteristics. In this dynamic environment, for a supplier of products and services to be successful it must be innovative and focusing on customer values and customer relationships. This proposition is based on the premise that customers are the most valuable assets of a company. Each transaction the company does with its customers is critical and must add value to the customer. Added values for the customers convert to added revenues for the company. The strategic question a successful company should ask is what innovative products and services it can offer its customers in order to provide them with solutions and interactions that enhance their values.

If understanding customer values is the key success factor, what about the competitors? In a dynamic environment where innovation counts more than productivity, competitive advantage is obtained by understanding the customers more than understanding the competitors. The reason is that a study of the competitors in an innovative environment is nothing more than a snapshot in time of a moving target. The study is outdated the moment it is ready. Therefore, a competitive positioning framework as we know it from, for instance Porter (1980) has less predictive value in this environment.

The consequences of the network economy on business development are outlined in the business model above (Figure 2). It includes new customer relationships, vertical segmentation of the value chain and horizontal aggregation of resources (content and customers). A strategy for a company entering into electronic commerce should be based on this model. We see three major aspects of a complete strategy for a company moving into the marketspace. Firstly, we advocate a stepwise procedure in developing a web site for a company. Secondly, a strategy must encompass ways of linking with business partners to enhance visibility and functionality. Thirdly, a company may take the opportunity to play a leading role as a web integrator (a business service provider).

Developing a Web Site

This strategy must be based on new opportunities of customer interactions offered by the Internet as outlined above. We advocate implementing this strategy in four steps.

Step 1: Create a web site (home page) to establish presence on the Internet. Present marketing oriented information on the company and its products and services.

Step 2: Establish transaction processing services integrated with back-end data bases and applications (e.g. order processing, bill payments).

Step 3: Interactive services with Java applications. Establish two-way interactive communication with customers. Products and services can be customized and proactive over the customers' life cycle; push technology can be utilized for updates, and customers may participate in product design and development.

Step 4: Develop horizontal aggregation facilities (e.g. brand communities) where customers can exchange experience with each other (FAQ, Bulletin boards, Newsgroups, etc.).

Interlinking Strategies

One of the key issues in an Internet strategy for electronic commerce is to obtain visibility. With millions of web sites residing on the net, how can potential customers find you? Existing, loyal customers will use the web address. Others may use a search engine. The problem with a search engine is to focus in on the right web sites in a search for potential suppliers of a particular product. The lack of good discriminating tools results in a search having too many outcomes, of which it is difficult to select the best. When the intelligent agent technology matures this situation may change. Today, however, unless you address the web site of the supplier directly, it is more by chance you may stumble over an adequate web site in your search.

One feature of the strategic importance of the Internet is the opportunity to move from one web page to another by hypertext links. These links enable users to move from web pages to web pages and from web sites to web sites, so-called *interlinking*. Interlinking is a means to establish presence on other web sites to guide visitors on these web sites to a company's home page. Interlinking can be used strategically as gateways or to build business partnerships (a

web within the World Wide Web). «Web-based strategies start from the assumption that the best way to manage risk and build capabilities and scale rapidly in highly uncertain environments is to mobilize the resources of many companies.» (Hagel et al., 1997). «Interlinking between business web sites is an important tool in the repertoire of the Internet strategist. Seamless linkages with business partners provide customers with more complete business solutions. Customers can navigate transparently between sites in ways that interlinking business partners, who have anticipated customers' buying habits and decision-making criteria, have established.» (Gascoyne and Ozcubukcu, 1996).

Interlinking is a means of web integration that can be organized by a business service provider, either on the basis of content (a content host), shopping (a shopping mall), or special interests (virtual communities). The business service provider can focus on particular customer segments or services (products, transactions), or both (broad based). To use some examples from Norway, *Bankplassen* (novit.no) provides links to all Internet banks in Norway and some foreign banks. It focuses on financial services. Another web site is *HA-Nett* (ha-nett.no) which has developed from a narrow customer segmented focus (activities of interest to people visiting or living in the county of Hedmark) to a broad based web integrator with links to local, national and international sites.

The strategic value of interlinking, however, hinges on two things: better visibility or more complete customer solutions. Better visibility is obtained by increased exposure, in particular if the host web is a frequently visited web site. A business service provider can shape more complete customer solutions. Linking up with a web site with a complete customer solution as a target can be very profitable to the interlinking companies. One may think of regional business service providers for real estate sales. The web site may assemble providers of services required to give customers a complete solution with respect to real estate purchases, like interlinking with real estate agencies in the region, financial institutions, government agencies, etc. It can even be extended to cover other services for people moving into the region such as car dealers, furniture shops or rental companies, etc.

Understanding the strategic value of interlinking is scarce and requires more research. There are, however, many examples to be studied.

Web Strategies for a Business Service Provider

A company may decide to take a leading role in augmenting its service offerings to the customers by creating a web host for other providers. This allows the company to deliver a broad range of services while preserving its own services for itself. Here are two examples taken from Hagel et al. (1997). Chase Chemical, a large US-bank, has opened a series of web sites that target particular segments among its customers. The Adventure Site, for instance, is a web site for those with an interest in recreational vehicles. On this site they can find information on trailer parks, share experience, etc.; and most importantly, they know where to go to get a loan if they are thinking of buying a new vehicle! The second example is a small regional bank in New England, Salem Five. This web site offers a wide range of services to customers interested in visiting New England.

1999 was the great year for new intermediaries in electronic markets changing the competitive conditions in almost any industrial sector of the economy. So far, however, the literature describing these online intermediaries is very fragmented and lacks a theoretical framework. In a paper by Methlie and Pedersen (2000), a model is presented that relates integration strategies of these intermediaries to structural and behavioral conditions of the marketspace. Four market condition components are identified: market, actor, product and integrated transactions. Thus the model is called MAP-IT, mapping market conditions into integration strategies. We will not develop this further in this paper. We can only refer to what is said about business service providers in the interlinking section above.

Conclusions

Markets play a central role in economic activity. Commerce is about the exchange of transactions related to the acquisition of products and services in markets. Internet as a platform of commerce changes market structures and the competitive conditions of the participants. Value chains are changed vertically upstream and downstream. Upstream value chains are disintegrated with more value creating activities exchanged in markets (outsourcing). Downstream distribution channels are disintermediated with vendors reaching customers directly without the use of traditional intermediaries such as agents or brokers, or they are reintermediated with new intermediaries, here called business service providers, entering the distribution channel. The

new market conditions require new business models and new business strategies. We have described how new customer relationships and new customer values can be created and how vertical and horizontal integration gives rise to new alliances.

References

Bailey, J P, Bakos, Y. An Exploratory Study of the Emerging Role of Electronic Intermediaries. *International Journal of Electronic Commerce,* 1 (3), 1997.

Evans, P B, Wurste, T S. Strategy and the new economics of information. *Harvard Business Review,* 71–82, 1997.

Dedman, R D. *Strategic Internet Provision : The emerging supply-side framework.* London, FT Media & Telecoms, 1997.

Downes, L, Mui, C. *Unleashing the Killer App.* Boston, Mass., Harvard Business School Press, 1998.

Gascoyne, R J, Ozcubukcu, K. *Corporate Internet Planning Guide.* N.Y., Van Nostrand Reinhold, 1997.

Hagel, J, Bergsma, E E, Dheer, S. Placing your bets on electronic networks. *The McKinsey Quarterly,* 2, 56–67, 1996.

Hagel, J, Armstrong A. *Net Gain.* Boston, Mass., Harvard Business Press, 1997.

Kelly, K. New Rules for the New Economy. *Wired*, 5.09, 1997. (http://www.wired.com/wired/5.09/newrules.html)

Katz, M, Shapiro, C. Network Externalities, Competition and Compatibilility. *American Economic Review,* 75, 424–440, 1985.

Learmonth, G P, Ives, B. Information System Technology can Improve Customer Service. *Data Base,* Winter, 6–10, 1987.

Malone, T W, Yates, J, Benjamin, R I. Electronic markets and electronic hierarchies. *Communications of the ACM,* 30 (6), 484–497, 1987.

Methlie, L B, Pedersen, P E. *MAP-IT : A Model of Intermediary Integration Strategies in Online Markets.* Paper submitted for the ICIS 2000 Conference in Brisbane, Australia, 2000.

Porter, M E. *Competitive Strategy : Techniques for Analyzing Industries and Competitors.* New York, The Free Press, 1980.

Porter, M E. *Competitive Advantage : Creating and Sustaining Superior Performance.* New York, The Free Press, 1985.

Rayport, J F, Sviokla, J J. Managing in the Marketspace. *Harvard Business Review,* November-December, 141–150, 1994.

Resnick, P, Zeckhauser, R, Avery, C. Roles of electronic brokers. In: *Toward a Competetive Telecommunication Industry : Selected Papers from the 1994 Telecommunication Policy Research Conference.* Brock, G W (ed.). Mahwah, NJ, Lawrence Erlbaum Associates, 289–306, 1995.

Stabell, C B, Fjeldstad, Ø D. Configurering value for competitive advantage : on chains, shops and networks. *Strategic Management Journal,* 19 (5), 413–437, 1998.

Thompson, J D. *Organizations in Action.* New York, McGraw-Hill, 1967.

Wigand, R T, Benjamin, R I. Electronic Commerce : Effects on Electronic Markets. *Journal of Computer Mediated Commerce,* 1 (3), 1995. (http://jcmc.huji.ac.il/vol1/issue3)

Winer, R S et al. *Choices in Computer-Mediated Environments.* University of California of Berkley, September, 1996. (Unpublished research paper.)

Quality Matters: Some Remarks on Internet Service Provisioning and Tariff Design*

JÖRN ALTMANN, BJÖRN RUPP AND PRAVIN VARAIYA

In order to ensure further Internet growth and efficiently support quality-differentiated network services, users' choice options have to go beyond different service plans that only reflect a rough market segmentation. To further subdivide these segments, choice options should provide the means to let users express their current needs by instantaneously selecting the service quality. This argument is strongly supported by empirical evidence from the Internet Demand Experiment (INDEX), a market trial for quality-differentiated Internet services. Arguing that the vast majority of Internet service plans for residential customers encourage waste and lead to user cross-subsidies, this article investigates user heterogeneity, activity heterogeneity, and acceptance of Quality of Service on demand to derive some consequences for Internet service provisioning and tariff design.

Jörn Altmann (34) is currently with the 'Internet and Mobile Systems Lab' of Hewlett-Packard Laboratories, conducting research on usage-based pricing of Internet services. Prior to this appointment, Jörn Altmann has been a senior scientist at the Department of Electrical Engineering and Computer Sciences at the University of California, Berkeley, and at the International Computer Science Institute at the University of California, Berkeley. Dr. Altmann received his B.Sc. degree in 1989, his M.Sc. in 1993, and his Ph.D. in 1996 from the University of Erlangen-Nürnberg, Germany. Dr. Altmann's current research centers around electronic commerce of network services.

altmann@eecs.berkeley.edu

Björn Rupp (28) is a Ph.D. candidate at Humboldt University Berlin and a member of the INDEX Project at the University of California, Berkeley. He is also affiliated with the Berlin-Brandenburg Graduate School of Distributed Information Systems. He received his B.Sc. degree in 1994 from the University of Heidelberg, and his M.Sc. in 1997 from Humboldt University Berlin. His research focuses on the economics of an Integrated Services Internet, spot markets for bandwidth, and the business impact of a high-speed Internet on telecommunications network operators and the media industry.

brupp@wiwi.hu-berlin.de

1 Background and Motivation

1.1 Quality and Waste

The economic model that the Internet has been using needs modifications to support new applications and different service qualities. However, current Internet service plans have made little progress towards supporting this goal: They are predominantly flat-rate and offer little, if any, choice of different service options. Currently, users who occasionally need large bandwidths are either forced to lease over-provisioned dedicated lines, risk the caprices of shared resources (best-effort quality), or forego the desired application altogether.

This leads to two kinds of inefficiencies: First, flat-rate pricing encourages waste. Whenever the marginal cost of network resource utilization is zero (like under a flat-rate pricing scheme), users do not have to optimize marginal utility and marginal cost. Inefficient over-utilization of resources occurs. Second, in a network with "best effort" shared resources like the current Internet, flat-rate pricing leads to noticeable quality deterioration for all users if those resources are over-utilized. Ultimately, if the network is consistently unable to fulfil the quality requirements that a certain application or individual user has, that user is effectively excluded from using the network for these purposes.

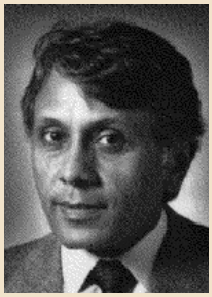
1.2 Integrated Services

Network externalities have fueled the trend towards an Integrated Services Internet. Different application classes (e.g. streaming video vs. file transfers) naturally have very different quality requirements in terms of packet loss, transmission delays, minimum required bit rate, etc. Users of different application classes also have heterogeneous Quality of Service (QoS) preferences.

Two basic network designs can meet those requirements: In case that connecting and subscribing to different networks is associated with very low costs, a market equilibrium with different network service providers offering different service qualities at different prices may prevail. With the emergence of a multitude of new services, however, it is at least questionable whether an increasing number of different networks tailored to support these services will not result in comparatively high costs. Subscribing to an integrated services network, in contrast, yields significant utility gains for every new network user. In addition, the potential penetration of new network services is increased significantly. New services can be accessed without the need for connecting and subscribing to another network, resulting in an incentive for the integration of as many services as possible in one network.

The integration of new services will ensure further Internet growth and allow for its even wider dissemination among the general population. The division of services into quality-differentiated market segments and the design of appropriate pricing structures for each segment are therefore crucial for further proliferation of Internet services.

* *This research was supported by grants from the National Science Foundation, Cisco Systems, SBC Communications, the California State MICRO Program, Hewlett-Packard and the German Research Society, Berlin-Brandenburg Graduate School of Distributed Information Systems (DFG grant no. GRK 316).*



Pravin Varaiya (60) is Nortel Networks Distinguished Professor in the Department of Electrical Engineering and Computer Sciences at the University of California, Berkeley. His areas of research are transportation systems, communication networks, hybrid systems, and electric power systems. From 1994 to 1997 he was Director of California PATH, a multi-university program of research in Intelligent Transportation Systems. From 1975 to 1992 he was also Professor of Economics at Berkeley. He has taught at MIT and the Federal University of Rio de Janeiro. He was a member of the technical staff at Bell Laboratories during 1962–1963. Dr. Varaiya has held a Guggenheim Fellowship and a Miller Research Professorship. He is a Fellow of IEEE, and a member of the National Academy of Engineering.

varaiya@eecs.berkeley.edu

1.3 Demand

In an increasingly competitive environment, service providers have to offer combinations of quality and price that match user needs. While much effort in recent literature has been invested in the design of adequate pricing proposals (for a short overview of different approaches, see (Shenker et al. 1996)), the understanding of the demand structure for quality-differentiated network services is still in its infancy. However, such understanding is critical for future network provisioning decisions.

2 Empirical Evidence

We aim to bridge the gap between supply-side and demand-side analysis by supporting the theoretical argument for flexible QoS choices with empirical evidence from the INternet Demand EXperiment (INDEX). INDEX is a real-world market trial for quality-differentiated network services. It provides Internet access over 128 kb/s ISDN lines to a group of users from the Berkeley campus community (students, faculty, staff). Users select network services from a menu of QoS-price offerings and pay for their usage. They control their usage of network resources by means of a Java application running on the user's computer (Figure 1). The subjects can choose a service quality instantaneously by clicking on a button and change the Quality of Service even during an active session. The application also provides usage feedback by displaying a summary of charges accumulated over the session, the day and the month. A detailed overview of the technology, experimental setup and design of INDEX can be found in (Rupp et al. 1998).

While we will occasionally refer to other INDEX sub-experiments for comparison purposes, the reasoning of this paper is mainly based on data from INDEX' first sub-experiment, *Variable Symmetric Bandwidth*, in which users are given the choice of six different bandwidths (8, 16, 32, 64, 96, and 128 kb/s). Subjects are charged a per-minute rate that depends on the selected connection speed. 8 kb/s service is priced at zero cents per minute. Prices for all other bandwidths are strictly increasing with bandwidth. They are randomly drawn from a set of prices ranging from a minimum of 0.1 cents/minute for 8 kb/s service to a theoretical maximum of 20.94 cents/minute for 128 kb/s service. Prices are varied during the experiment to measure the demand response for each individual. The data in this paper was collected by analyzing the usage patterns of 70 subjects. It covers the period from April 1998 to February 1999.

2.1 User Heterogeneity

The utility that is derived from the consumption of network resources depends on individual characteristics and preferences. This subsection highlights the extent of user heterogeneity by examining the budget variation for Internet usage, time spent on the Internet, and the weekly mean expenditure.

Figure 2 shows a histogram of the weekly mean expenditure of the subjects over three different experiments. Besides data from the *Variable Symmetric Bandwidth* experiment, we also use data from the *Variable Asymmetric Bandwidt*h experiment (in which subjects can choose differ-

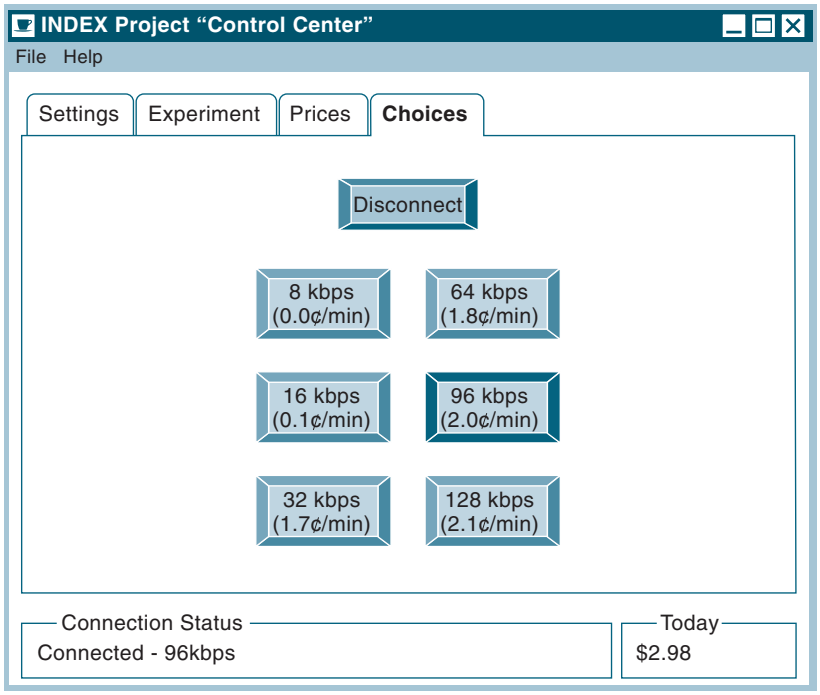


Figure 1 INDEX User Interface

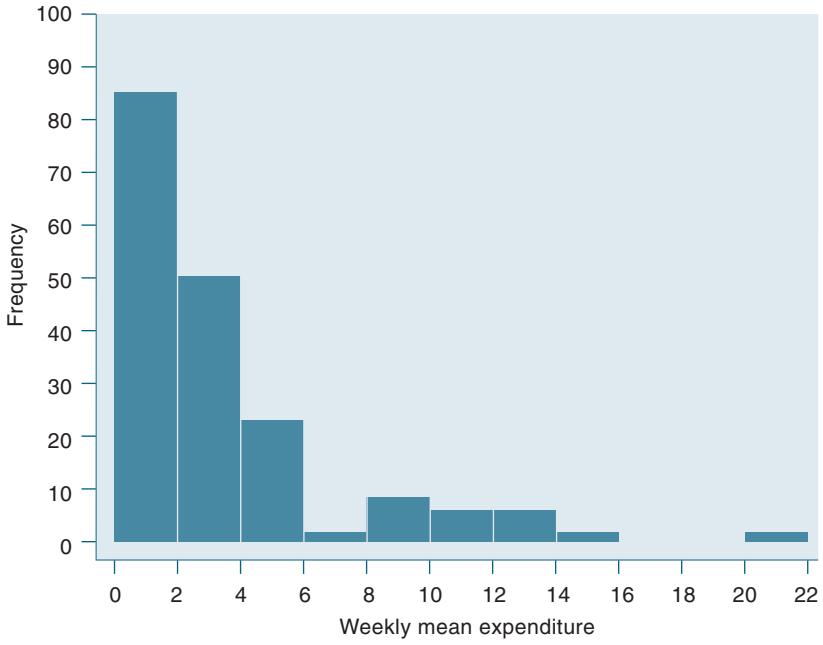
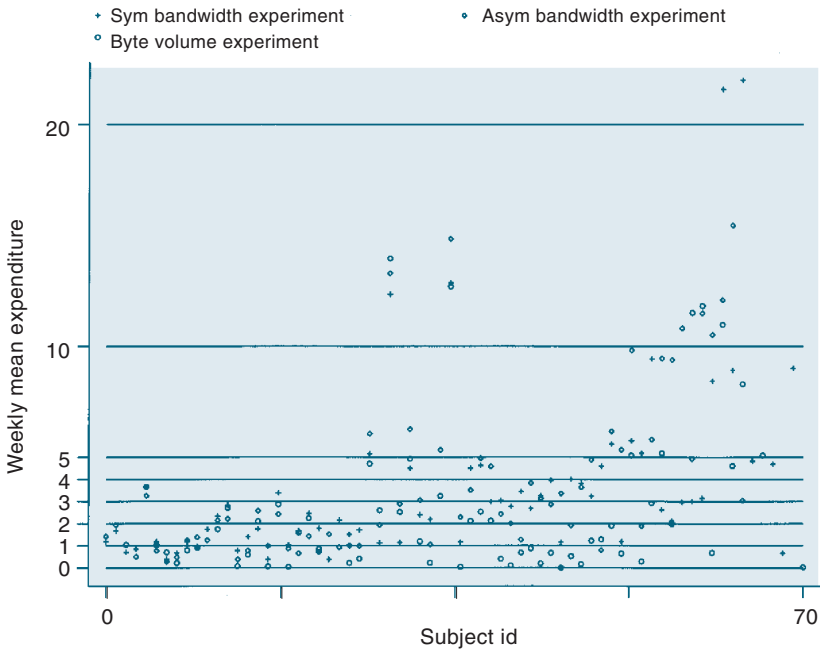


Figure 2 Expenditure Histogram

ent bandwidths for traffic *from* the Internet and *to* the Internet separately) and the *Byte Volume* experiment (in which subjects pay for the number of transmitted bytes). That compensates for the effect of different pricing structures. Therefore, each subject is represented three times, once per experiment. The variation in subjects' weekly mean expenditure spans a range from $ 0.20 to $ 21.23. Although about 40 % of the subjects spend less than $ 2 per week, there is a considerable variation in expenditure in the $ 2 – $ 22 range.

Instead of charging for individual usage, we could impose a flat-rate tariff set to recover the same revenues (by dividing total expenditure by the number of users). In this case, we would however only address the needs of 25 % of the subjects, i.e. people spending between $ 2 and $ 4. We would deprive ourselves of the additional revenue that could be generated by the users with a higher willingness to pay. At the same time, we would exclude users with only very casual Internet activity from using the service at all.

Figure 3 Budget



Looking at the differences between the three experiments, it becomes obvious that the disparities in expenditure patterns are not simply due to environmental or seasonal effects. They are rather an inherent characteristic of individual demand. Figure 3 shows this effect: The mean weekly expenditure of 48 % of all INDEX subjects varied only in a range of $ 2. Considering that users were facing widely varying prices and disparate pricing schemes over the course of the three experiments, this result gives clear evidence that a significant percentage of users does have an exact idea of how much they intend to spend for Internet services in a given time period. Another 24 % of our subjects also set relatively tight constraints on their Internet budget, allowing for a maximum variation of $ 4. Only the remaining 26 % displayed a significantly wider variation in their expenditure distribution.

Beside user expenditure, the time that users spend online is another indicator of user heterogeneity. Figure 4 compares per-minute pricing with flat-rate pricing. It demonstrates the high degree of cross-subsidy between light and heavy users under a flat-rate tariff. To analyze this effect, we aggregate and normalize the expenditure data on a per-user basis. We then rank the 70 subjects by the time they spent online (processing heaviest users first). The resulting upper curve of Figure 4 plots the actual cumulative expenditure from the *Variable Symmetric Bandwidth* experiment versus cumulative connect time. The curve starts out from the heaviest users close to the graph's origin and proceeds to the lightest users on the far right. Each dot represents one subject.

Each user is then imputed a flat-rate expenditure. Under this flat-rate pricing scheme, the imputed flat rate times the number of users equals the total revenue generated by the actual per-minute charges from the *Variable Symmetric Bandwidth* experiment. The lower curve in Figure 4 represents the cumulative expenditure under this flat-rate tariff.

The calculations yield results that relate well to other observations from the telecommunications field. 20 of the 70 subjects (28.6 % of the subject population) consume more than 75 % of network

resources, measured in connect time to the Internet. Under flat-rate pricing, these heavy users would be subsidized by light users and would account for only about 30 % of overall expenditure. However, under per-minute pricing, these 20 users are charged in proportion to their usage. They have spent approximately 45 % of the actual overall expenditure. This demonstrates that usage-based pricing is a fair way to charge people and is significantly more equitable than the predominant flat-rate Internet service plans.

2.2 Activity Heterogeneity

After exposing the high degree of inter-user heterogeneity, we now turn to intra-user heterogeneity. In particular, we will look at two factors that show how individuals use the Internet for different purposes: time of day and activity type.

Classical peak-load pricing models heavily rely on time of day for allocating resources. Such pricing models assume that demand for a non-storable good such as bandwidth can be divided in different sub-periods in which the demand function does not change significantly. While we still see notable demand variations in terms of QoS choices made during a given period, we have been able to identify temporal patterns in network resource consumption. Figure 5 shows that network usage changes over the course of a day. Time of day (measured in minutes) is plotted versus frequency of use. One dot represents measured activity at the minute indicated, accumulated over the entire 6-week duration of the experiment. The underlying data is again taken from the *Variable Symmetric Bandwidth* experiment. The lower curve represents usage of paid services (16 kb/s, 32 kb/s, 64 kb/s, 96 kb/s, 128 kb/s). The upper curve represents total usage (i.e. paid service plus free 8 kb/s service).

We can easily identify times of high traffic density, with a clearly visible peak time from 8 to 9 p.m. versus very little traffic in the early morning hours around 5 a.m. These findings are consistent with other data gathered on UC Berkeley modem pool usage.

While this data suggests that certain temporal patterns of residential Internet usage are relatively predictable, the detailed actual usage patterns show a wide variation in type and extent of usage. The same network infrastructure is used for very different tasks. Different applications and preferences lead to heterogeneity in individual network usage and resource consumption.

To illustrate this, we will focus on three different types of activity: *Bulk Traffic* (e.g. FTP, streaming data), *Burst Traffic* (e.g. World Wide Web), and *Interactive Traffic* (e.g. Telnet, X Windows).

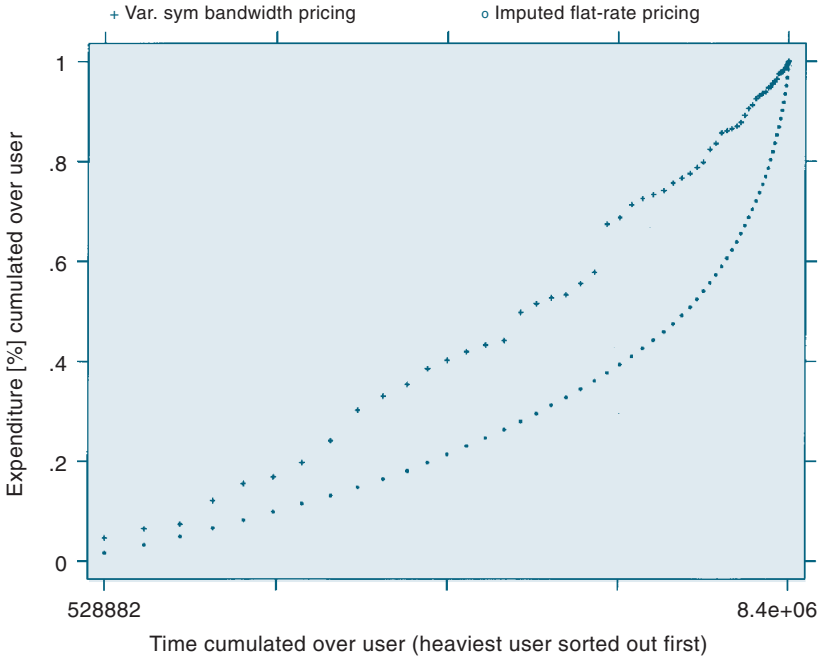
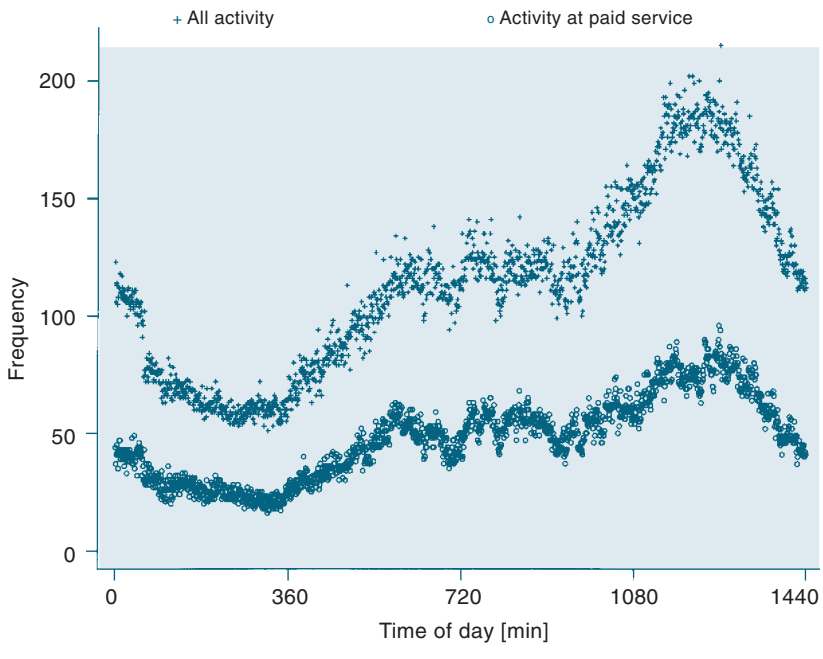


Figure 4 User Cross-Subsidy

In order to determine the activity type, we measure the average size of packets transferred during a minute. The classification is derived from this measure of traffic density and carried out for each minute of recorded in-bound traffic. As a consequence, we can only classify based on the main activity type in a given minute (i.e. either bulk, burst, or interactive traffic, but not a combination of different activity types). Activity is defined as bulk traffic if the number of packets is small compared to the number of bytes transferred, i.e. the average packet size is larger than 1000 bytes. If the average packet size is smaller than 45 bytes, then we infer that that minute was mainly used for interactive applications. All

Figure 5 Time of Day Activity



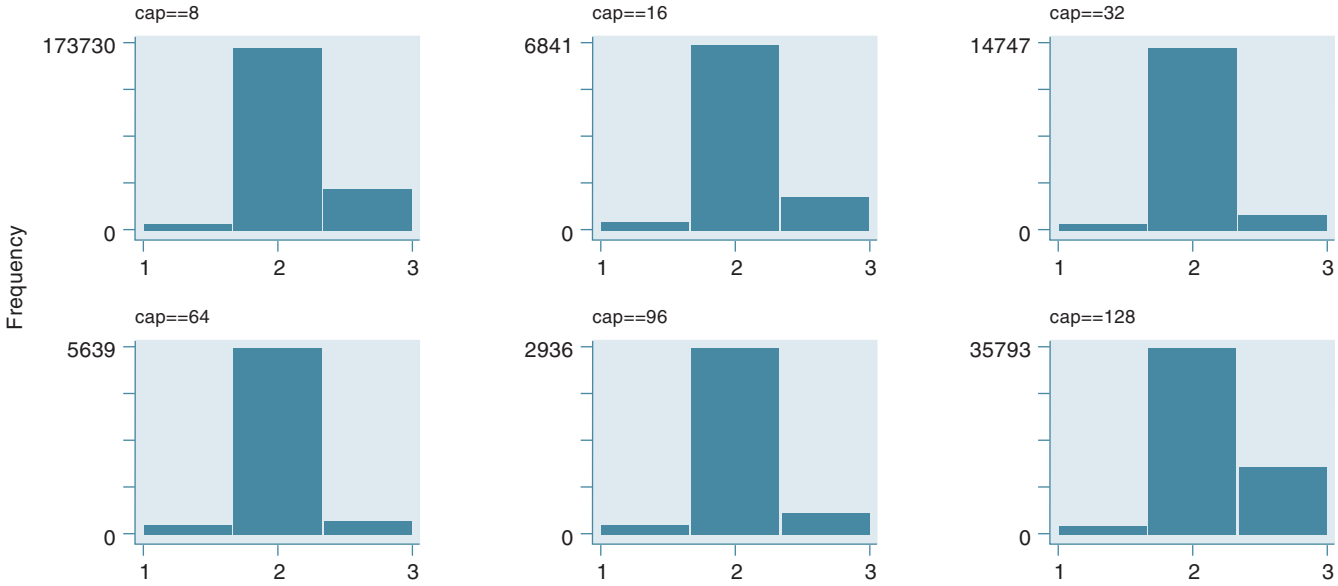


Figure 6 Activity Type

remaining minutes of recorded traffic are classified as representing bursty traffic. This classification method is of course limited insofar as it is based on proxies. Nevertheless, it can be used to point out some interesting properties of usage patterns.

Figure 6 shows the distribution of the three activity types for each choice of bandwidth (8 to 128 kb/s). The activity type is represented by a number: 1 stands for bulk traffic, 2 for burst traffic, and 3 for interactive traffic. The overall percentage of burst traffic ranges from 66 % to 85 %, depending on bandwidth choice. Burst traffic clearly is the dominant activity type. An interesting observation is that interactive traffic strongly increases at 128 kb/s. This illustrates that the subjects value fast response times although they are effectively wasting purchased capacity. These results verify the findings of a survey of residential Internet users in the mid-western USA about high-speed Internet access (Hoag 1997).

3 Acceptance of QoS on Demand

After illustrating the heterogeneity of users and activities in the previous section, we now examine whether users appreciate the higher flexibility of being able to make QoS decisions at any time.

Figure 7 QoS Choices – Per Subject

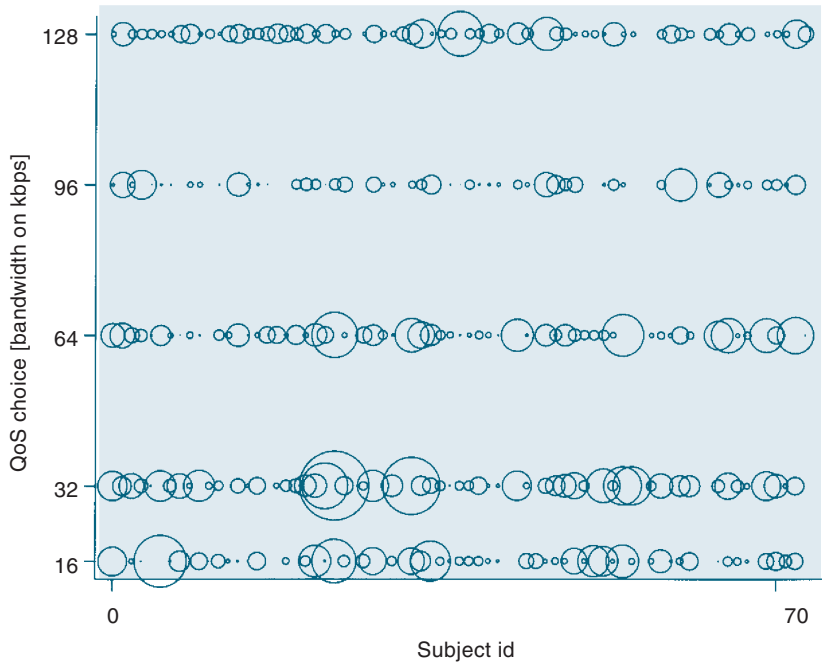


Figure 7 depicts the QoS choices of all INDEX subjects. The weighted two-way scatter plot graphs each subject against all priced bandwidth choices made by that subject over the entire duration of the *Variable Symmetric Bandwidth* experiment. The sum of all minutes spent at the same bandwidth is graphed as one circle, with the radius of each circle being proportional to the share of total time spent online by that subject and at that bandwidth.

The diversity of QoS choices pictured in this graph visualizes two important characteristics of user behavior: First, different users obviously have different valuations of network resources and consequently prefer different bandwidths. Second, QoS choices of individual users are not restricted to just one or two bandwidths. The subjects make use of a wide range of bandwidths. Almost all users purchase high quality

at least sometimes. Figure 8 emphasizes this second characteristic even more. The histogram displays the number of different QoS choices (priced and unpriced) that were made over the entire duration of the experiment: 62.5 % of the subject population made use of the entire range of options (8 kb/s to 128 kb/s). 75 % used at least five different bandwidths.

These results make a strong case for persistently giving Internet users the choice of multiple service qualities. Our subjects clearly endorsed the higher flexibility associated with being able to make QoS decisions at any time. If different service qualities are only offered by means of subscribing to them under corresponding static flat-rate pricing regimes, then users who only occasionally demand high-quality services will be excluded from using such services. In contrast, if charges are set in proportion to actual usage, high-speed services are accessible for a much broader set of users. INDEX data suggests that the range of different quality choices will be fully exploited if the economic incentives to make use of these choices are not distorted by static service plans.

4 Implications for Service Provision and Tariff Design

The preceding analysis has shown: There is little intra-user variation in weekly mean expenditure; the inter-user distribution of individual budgets for Internet access and general usage intensity is very diverse; demand varies not only over time, but also depends on the activity type.

These findings lead us to conclude that user heterogeneity and activity heterogeneity call for a flexible system which enables users to choose Quality of Service on demand. Price discrimination that is based exclusively on customer type (i.e. different service plans for high-volume and low-volume users) is not sufficient to meet the demand of tomorrow's Internet users. In contrast, users should also be given the opportunity to switch between different service qualities more or less instantaneously. We have shown that users endorsed the higher flexibility associated with being able to make QoS decisions at any time. A wide range of QoS choices were made. In the *Variable Symmetric Bandwidth* experiment that we analyzed, 62.5 % of the subjects made use of *all* the QoS options that were offered to them.

At a general level, QoS on demand avoids waste and user cross-subsidy. Quality differentiation, combined with proper economic incentives, increases the overall value of a network and can pave the way towards an economically viable Integrated Services Internet.

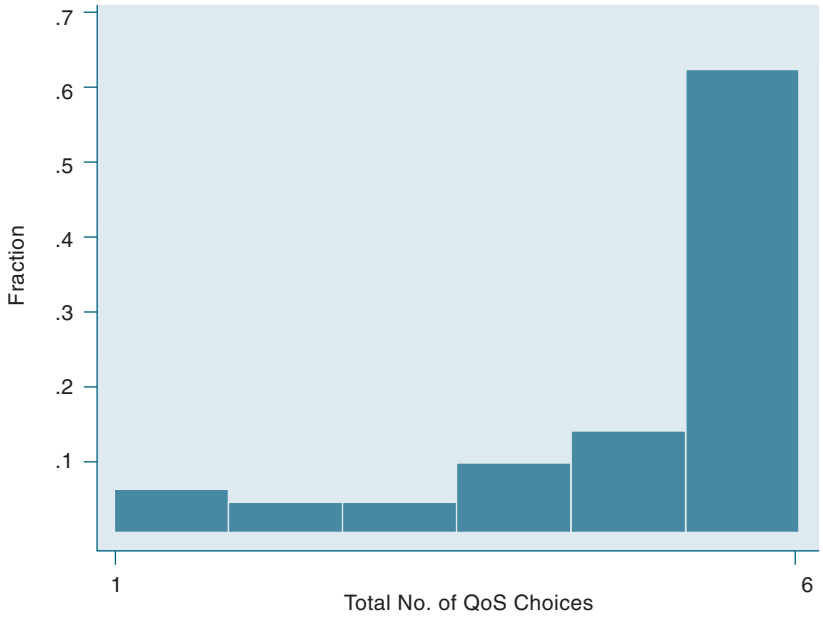


Figure 8 QoS Choices – Histogram

References

1 Hoag, A. 1997. Speed and the Internet : The effects of high speed access on household usage. In: *Proceedings of the 25th Telecommunications Policy Research Conference,* section IX, 24–37.

2 Rupp, B et al. 1998. INDEX : A platform for determining how people value the quality of their Internet access. In: *Proceedings of the 6th IEEE/IFIP International Workshop on Quality of Service,* 85–90.

3 Shenker, S et al. 1996. Pricing in computer networks : Reshaping the research agenda. *Telecommunications Policy,* 20 (3), 183–202.

Interconnection and Competition Between Portals Offering Broadband Access*)

ØYSTEIN FOROS AND BJØRN HANSEN

In this paper we consider competition between two broadband access providers offering differentiated portal services in the same geographical market. Our analysis indicates that in cases where the ability to attract broadband content is an increasing function of network size, and one expects that the outcome of the market game is characterized by more than one firm being active, then the strategy of degrading or choosing low interconnection quality is sub-optimal (a so-called hedged garden strategy). In a duopoly, both firms will be better off by choosing high interconnection quality. As demonstrated in the paper, this conclusion also holds in an asymmetric equilibrium. The analysis in the present paper may be of particular relevance in cases where a cable-tv operator and a telephony operator upgrade their access networks in order to offer competing broadband portals/services.

Øystein Foros (31) holds a Master of Business and Economics degree (1993) from the Norwegian School of Management (BI) and is Cand. oecon. (1996) from the Norwegian School of Economics and Business Administration (NHH). He has been employed as Research Scientist at Telenor R&D since 1996. Since 1998 he also holds a position as Scientific Advisor to the Foundation for Research in Economics and Business Administration (SNF). Current areas of specialization are industrial organization analysis of the Internet and telecommunication markets.

oystein.foros@telenor.com

Bjørn Hansen (34) is Research Manager, Network Economics and Strategy at Telenor R&D. He is Cand.Oecon. from the University of Oslo (1992). Before joining Telenor he was a research associate at The Foundation for Research in Economics and Business administration at the University of Oslo. Current research interests are competition and business strategy in the converging Information, Computer and Telecom sectors.

bjorn.hansen@telenor.com

I Introduction

Consumers will not subscribe to broadband services unless both the necessary access bandwidth is available and interesting information goods are delivered over this bandwidth, ie. access bandwidth and information goods are strict complements. This may lead to a nontrivial problem of co-ordination. In this paper we consider competition between providers of access bandwidth that seek to solve this co-ordination problem by taking an active role in making broadband information goods available to their customers. The broadband access providers sign contracts with owners of copyrights to broadband information goods such that they become available within the network domain of the access provider. Thus the access providers take a business role as access portals.[1)] We assume that the ability to attract content by the access portals is an increasing function of the number of subscribers.

There are several technological approaches for connecting residential customers for high-speed data communication, what we call broadband access. First, to build an entirely new fiber-optic line to customers' thresholds (fiber-to-the-home) gives superior speed of communication compared to upgrading existing lines into homes. However, the high up-front investments of new wireline facilities imply that there will be no rush to install additional wires to homes (see Clark, 1999a). In a given area the most likely providers of broadband access will be the telephone operator controlling the copper lines and the cable-tv operator controlling the coaxial cables. Upgrading existing lines is much cheaper than installing fibers to each subscriber. By upgrading their networks they can use the existing copper lines and coaxial cables the last path into homes. Cable modems can be used in the cable-tv network, while the capacity in the copper loop can be increased by DSL-technology. The longer the distance the copper/coaxial cable is used for, the lower the capacity.[2)] Due to the cost advantage described above we will concentrate the analysis to the case where a cable-tv operator and a telephony operator upgrade their access networks in order to offer competing broadband access.

At an equal price all customers will prefer the fiber-to-home-solution to the upgrading solution. In contrast, there is no simple metric to rank between the broadband access service over the copper line and the coaxial cable (see Clark, 1999b). As is well known, there is a significant cost of upgrading a customer to broadband, and firms with some installed infrastructure will have considerable advantages since they already have incurred a portion of this cost. The cost of upgrading a single customer will however depend upon properties of the installed infrastructure as well as some other factors. The cost factors can for instance be local loop length and quality of transmission facilities. Since the exist-

*) *This work is partly financed by the IST programme.*

1) *The term "portal" is currently being used to describe almost any type of business on the net. Here it is used in a quite narrow and specific sense. A broadband access portal is a provider of broadband connectivity to the (Inter)net that seeks to make broadband information goods (content) available to its customers.*

2) *For example for copper lines 2 Mbit/s symmetric capacity can be delivered over a cable up to 2.5 km, whereas 26 Mbit/s symmetric capacity can be delivered over a cable of only up to 0.3 km (see eg. Ims, 1999).*

ing networks have been built at different points in time for different purposes (telephony, tv) the networks will differ with respect to the properties of the installed infrastructure. Thus, we assume that the services offered by the telephony provider and the cable-tv provider are horizontally differentiated as seen from the customer's point of view.[3)]

A critical decision for the market players in such a context is the quality of interconnection between the networks, ie. to what extent is it profitable for provider *A* to choose the same quality for on-net and off-net communication. Off-net communication can be a video-conference between subscribers connected to different providers, or it may be a subscriber of provider *A* accessing content offered by provider *B* and *vice versa*. For conventional Internet services, such as email and www, the strategic importance of interconnection between competing access providers has not been given much attention. There may be several reasons for this.

First, the quality of local interconnection between competing providers in the same area was rather unimportant since the majority of Internet content was located in the US. As broadband is introduced, the situation can be expected to change since a cost effective way of distributing information goods (from the US or anywhere else) is to establish local portals where heavy content is downloaded such that one can economize on capacity in expensive long distance links. In particular this will be important for broadband providers outside the US. Thus the portion of the Internet traffic where both the customer and the server are located in the same area can be expected to increase (see also Mueller et al. 1997). This tendency will be reinforced by new customer types and new services in the Internet. For example in non-English speaking countries content intended for the mass market must be produced locally or translated (or provided in different versions for each sub-market). Furthermore, for interactive communication services (eg. interactive videoconferences) the portion of local traffic will probably be higher than for services like www.

Second, the *best effort* standard in the Internet ensures a sufficient "interconnection" quality for services like email and www. In contrast, broadband services require a higher level of quality that the competing providers must ensure through interconnection agreements. A firm like AOL/Time Warner, for instance, may find it profitable to give access to some broadband services exclusively for its own customers. In such a context the market share of the provider may be of more crucial importance for the customers for broadband services compared to conventional Internet services (see Shapiro and Varian, 1998).

Taking these features into consideration, it seems likely that the quality of local interconnection will be more important seen from the consumer's point of view, and accordingly, the importance of local interconnection as a strategic variable has increased.

Broadband customers are evidently subscribing in order to get access to broadband content (audio, videos, games, etc.). The broadband access providers will accordingly have to sign contracts with firms that have copyright to such content. The ability of an access provider to attract broadband content will typically be an increasing function of the number of customers connected to the respective access provider. The customer willingness to pay for subscription will increase in the variety of available content, and thus there are positive indirect consumption externalities. As the number of subscribers of an access provider increases, the variety of available broadband content will increase and thus the willingness to pay for subscription will increase. This effect is a positive consumption externality. Broadband is also expected to lead to more direct customer-to-customer communication such as new interactive services (telemedicine, tele-education, videoconferencing, etc.). This development will further reinforce the significance of local interconnection. Furthermore, a consequence of willingness to pay increasing in available content is that by increasing the interconnection quality, subscribers will get access to off-net content. Thus, for given market shares, the customer's utility and accordingly, their willingness to pay for subscription, will increase in interconnection quality. It is not obvious, however, that competing firms will choose a high quality. In the presence of network externalities, customers will *ceteris paribus* consider it more advantageous to choose the larger provider if the chosen quality of interconnection is reduced. A large provider may accordingly choose a low interconnection quality in order to increase its market share. In this paper we focus on effects from client server traffic as broadband is introduced.

[3)] *Several recent studies of the competition in the telecommunication market, eg. Laffont, Rey and Tirole (1998a, 1998b), assume that firms offer horizontally differentiated goods. The motivation for this horizontal differentiation is receiving little attention in the literature. In our setting, however, product differentiation in the horizontal dimension may be given a concrete interpretation.*

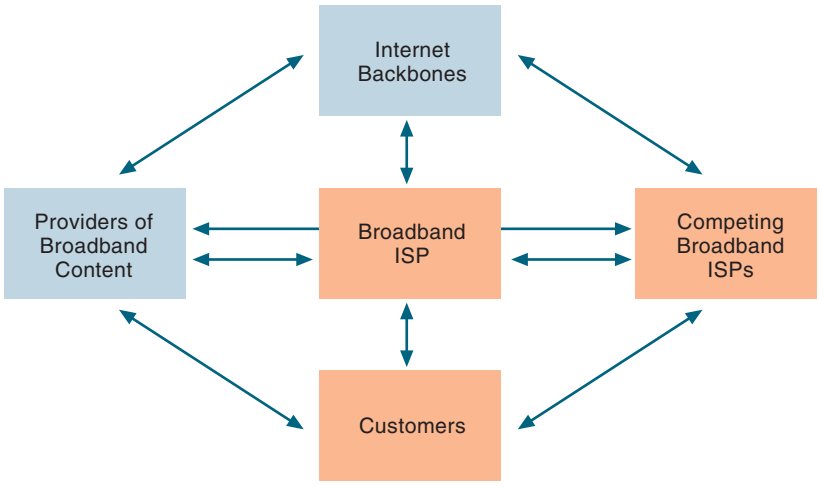


Figure 1 The business environment of a broadband ISP

In Section II we present a stylized business model, and discuss some key features of the market for broadband access to residential customers. In particular we describe how the firms can affect the interconnect quality. In Section III we analyze the business model in a duopoly context with horizontal product differentiation and price competition. We use the model of Foros and Hansen (1999a, 1999b) to analyse the choice of interconnection quality. We model the competition between the two broadband ISPs (Internet Service providers) as a two stage game where the firms in the first stage determine the quality of interconnection (ie. to what extent, or at what quality, customers get access to broadband content from the other firm). This choice of interconnection quality can be considered as a choice of interconnection quality between the networks. In the second stage, for given interconnection quality, the two firms compete à la Hotelling in attracting customers.

Our analysis indicates that in cases where the ability to attract broadband content is an increasing function of network size, and one expects that the outcome of the market game is characterized by more than one firm being active, then the strategy of degrading, or choosing low interconnection quality is sub-optimal (a so-called hedged garden strategy). In a duopoly, both firms will be better off choosing high interconnection quality. Finally, in Section V we conclude.

For more references to the specific theoretical literature, see Foros and Hansen (1999a, 1999b). For an extensive discussion of interconnection strategy related to telecommunication firms, see Laffont and Tirole (2000).

II A Stylized Business Model

Since the broadband ISP business is in its infancy it is far from evident what particular business model one should consider. In this section we will describe a stylized business model that will serve as the object of strategic analysis in the following sections. The broadband ISP under consideration is providing Internet access to end-customers charged on a flat rate basis (ie. the customers pay a fixed fee per month). Our focus is on competition in the end-user market and the choice of interconnection quality, and we will accordingly model these aspects explicitly. A broadband ISP must however also have some sort of business relationships with Internet backbone providers and providers of broadband content.[4)] Providers of broadband content is eg. owners of copyrights to music and films as well as application service providers (ASPs; entities offering downloadable software over the network such as games, business software, etc.). In figure 1 we have illustrated the business environment of a broadband ISP.

For a user of the Internet, connectivity and content are strict complements; ie. connectivity is without value unless content is available and *vice versa*. A narrowband ISP however does only to a limited degree have to take this complementarity into consideration since connectivity to the global Internet will give the users access to vast amounts of narrowband content anyway. This is in contrast to a broadband ISP. In general it is both expensive and technically challenging to offer connectivity at sufficient (guaranteed) bandwidth to the global Internet, eg. connections to the US is by expensive private leased circuits.[5)] Once customers have broadband accesses the importance of this cost effect will increase since the required bandwidth increases considerably. Upgrading a customer from ISDN speed to say 380 kbit/s yields an increase by 6 in required bandwidth, whereas if users want an access bandwidth able to carry two High Definition tv-channels, the required bandwidth increases by a factor of 235. This cost effect is particularly important for ISPs outside the US since, as argued above, the cost of providing bandwidth to the US is considerable.[6)]

[4)] *By the terms* broadband *and* narrowband *content we have in mind content that require broadband and narrowband connectivity respectively.*

[5)] *Baake and Wichmann (1998) and Mueller et al. (1997) emphasize that the cost-reduction from setting up a local interconnection point, instead of sending local traffic via USA, may be substantial.*

[6)] *This cost is increasing in distance to the US and the cost is accordingly significant in Australia, see eg. Ergas (2000) and Little et al. (2000).*

A cost effective solution is accordingly that broadband content is downloaded to servers within the domain of the broadband ISP. In this way the broadband ISP can offer customers service guarantees with respect to the quality of service.[7] If a broadband ISP fails both to attract broadband content and to offer broadband connectivity to the global Internet, the broadband aspect of the service is of no value to the users since content only is available at narrowband quality. Thus a successful broadband business strategy has to include a proactive strategy towards content providers.

In the business model considered here, the content providers establish servers within the domain of the broadband ISP. The content provider is then establishing a direct relationship to the end-customer such that the content provider can charge the end-customer in the way it finds suitable. One example can be an on-line video library where end-customers via their broadband access can download movies. The movies can eg. be sold at a unit price or the movie library can derive its revenues from advertisements. Similarly an ASP can charge its customers by usage or by a monthly fee.

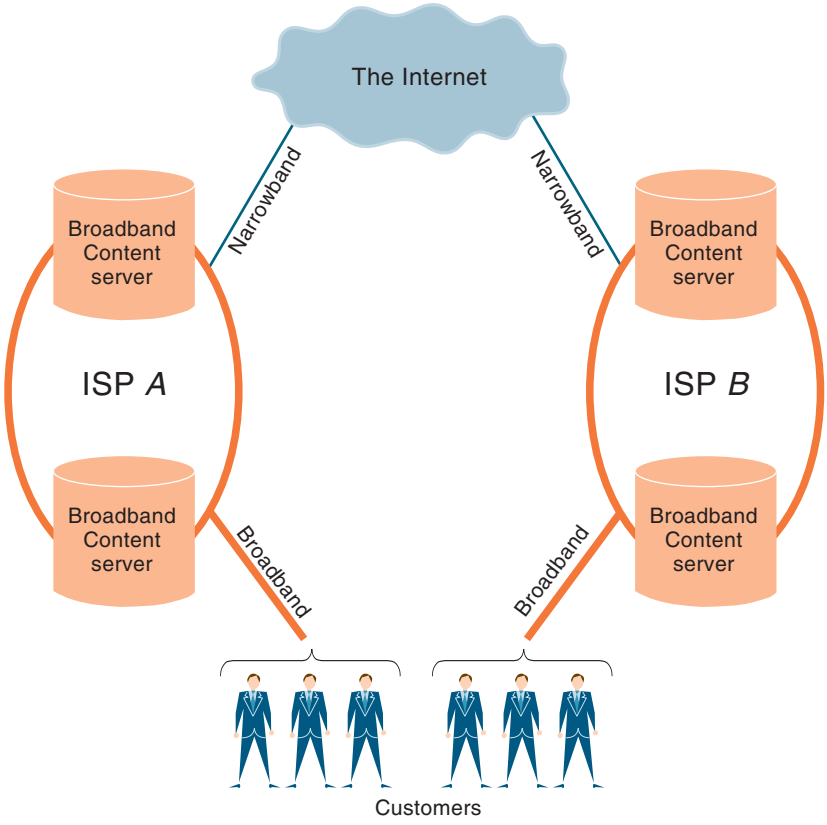


Figure 2 "A hedged garden environment" without direct interconnect

In this paper we are not focusing on the interaction between the broadband ISPs and the backbone providers and content providers. Furthermore, we are not looking into the different business strategies a content provider will choose. We are modelling these aspects in a passive way.[8] To emphasize this we have given a blue shade to the backbone and the content roles in figure 1. As far as the relation to Internet backbone providers is concerned we will only assume that the local broadband ISP is connected to the global Internet at narrowband quality. Furthermore, we are including a positive feedback loop into our model – by attracting more customers, the broadband ISP will attract more content that will attract more customers, etc. As argued earlier, this is an example of positive indirect network externalities.

In figure 2 we provide an illustration of how the different entities are linked together.

"Traffic" or communication between customers and servers located on the same network is called on-net traffic, whereas traffic between a customer of network *A* and a server in network *B* is called off-net traffic. ISP *A* is offering a quality guarantee of $\overline{k}$. If ISP *A*'s customers are communicating with a server located on the network of ISP B (off-net traffic), no such quality guarantee is given. Off-net traffic is assumed to be narrowband and the quality is $\underline{k}$ (where $\underline{k} \leq \overline{k}$). Such a market structure is by some denoted a hedged garden strategy since customers and content within the same domain are of high quality whereas across domain, or across the hedge, connections are of low quality.

As indicated in the text above, the indirect network externalities yield a positive feedback loop such that the willingness to pay for broadband ISP connection for the customers increases with available broadband content. Thus the two ISPs may install a direct broadband link between the two networks such that the total available broadband content increases for customers of both ISPs. Installing such a high capacity broadband interconnection link will result in two opposing effects. On the one hand, the willingness to pay for broadband access increases and thus the profit potential for both ISPs. On the other hand, installing such a link may affect the competition in the end-user market. In particular, by installing a high capacity interconnection link, an ISP will make the competing firm more competitive since the value offered to subscribers in-

[7] *The guarantee can be with respect to effective bandwith, delays, etc.*

[8] *This is an object for further study since there is a set of analytically challenging questions to consider by modelling the interaction between content and connectivity providers respectively.*

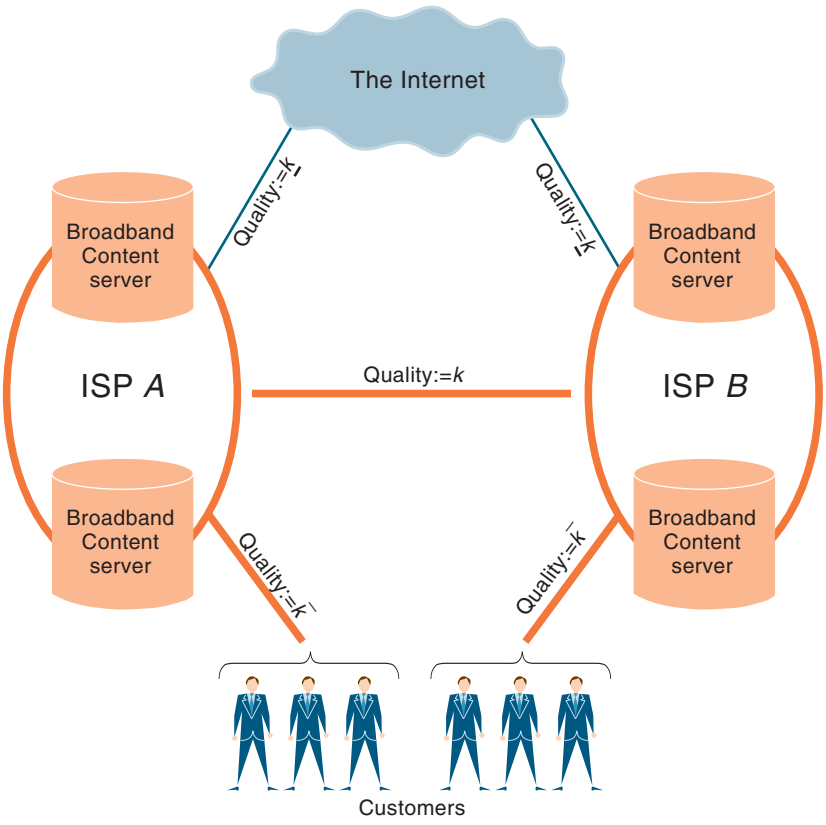


Figure 3 An "open environment" with interconnection

creases. Thus the installation of a broadband link may result in reduced market shares for a broadband ISP.

If the two competing ISPs agree upon direct high capacity interconnection they can install a high capacity link with quality k, where $\underline{k} \leq k \leq \overline{k}$. Such a link will ensure better quality of the communication between the competing ISPs. Figure 3 illustrates the context when there is a direct interconnection link between ISP *A* and ISP *B*.

The routing structure in this diagram (figure 3) is the same as in figure 2, with one exception. ISP *A* and ISP *B* have invested in a direct interconnection link which serves traffic between ISP *A*'s customers and ISP *B*'s servers and *vice versa*. The quality level of this traffic is in the interval $[\underline{k},\overline{k}]$. The aim of this paper is to analyze the incentives competing broadband ISPs have to implement such direct interconnection.

In the ISP markets in the European countries competing narrowband ISPs have been reluctant to implement direct interconnection links between their networks in order to enhance the quality of off-net traffic to ISPs in the same area. Thus, they operate in a context more like figure 2 than figure 3.[9)]

We have used the quality-term in a specific way (bandwidth). However, the quality of interconnection could be given several interpretations (see Foros and Hansen, 1999a and 1999b). As mentioned above, the key notion is that customers have preferences for high quality and that quality degradation of interconnection prevents them from communications with off-net servers with the same quality as communications with on-net servers. Note also that the motivation for degrading the quality of an interconnect arrangement in the ISP market may be analogous to the price discrimination between on-net and off-net traffic we particularly observe in the mobile telephony market. In the mobile market, we see that operators charge different prices for calls terminating on-net and calls terminating on a competing network (off-net). Price discrimination between on-net and off-net calls may squeeze an operator with a small market share and create a *de facto* degradation of interconnection.

III The Market Model

In this section we will provide a detailed and analytical discussion of the market interaction that will result from a game between two broadband ISPs employing the business model described above. The technical part of the discussion is taken from Foros and Hansen (1999a), but we will try to put emphasis on the link between the presented analytical model and the business model discussed in the previous section.

The Demand Side

In general the potential customers of the broadband ISPs are identical in some dimensions and they differ in other dimensions.

In the previous sections we have argued that the cost of upgrading a customer to broadband differs due to properties of the installed infrastructure such as local loop length and quality of transmission facilities. Since the existing networks have been built at different points in time

---

9) *Take Norway as an example: The dominating ISPs are managing their own networks and off-net traffic between the competing ISPs is routed through the Norwegian Internet eXchange (NIX). The NIX is non-commercial and managed by a third party administrator at the University of Oslo. It has been argued that the NIX constitutes a bottleneck even for narrowband services.*

10) *In our analysis we focus on the competition between cable- and telephony-based players. We would obtain similar results if we investigated duopoly competition between any two players based on LMDS (fixed wireless), UMTS (3. Generation mobile), etc.*

for different purposes (telephony, cable-tv) the networks will differ with respect to the properties of the installed infrastructure. Thus the two firms will be horizontally differentiated since one of the firms has an advantage serving some customers, and the other firm has an advantage serving other customers.[10)]

In addition to the cost of infrastructure upgrade source of product differentiation there is a "usage pattern" source of product differentiation between broadband ISPs based on telephony and the cable-tv business respectively. In broad terms a cable-tv-based broadband ISP can be expected to have an advantage in serving "entertainment"-heavy customers whereas the telephony based broadband ISP typically will have an advantage in serving home offices and more professional needs. The source of this difference is partly technical. A cable-based ISP will have an infrastructure suitable for broadcasting and entertainment-like applications whereas the telephony-based ISP has an advantage in providing guaranteed point-to-point connections. The other part of source for the entertainment – professional differentiation is the simple fact that the rooms where customers typically have their telephony and tv jacks respectively differ. A consumer will have to incur a cost in order to get a connection from the cable-tv jack in the living room to the pc in the home office. Similarly, it is costly to connect the telephony jack in the home office (or the hall) to the entertainment center in the living room.

We will assume that all horizontal product differentiation due to the factors described above, can be described as a cost term specific to each consumer. This will be formalized by assuming that each consumer has preferences in the horizontal dimension (cost of upgrade, usage pattern) that can be described by a location on the unit line. The services offered by the two suppliers are located at the extremes of this unit line. Thus consumers with a location of preferences close to service offer i will tend to choose this service unless the price of the other service is sufficiently low to compensate for the difference between preferences and product type. When a customer with preferences located at x chooses to connect to a supplier located at x_i ($i = a, b$) she will incur a "transportation" cost equal to $t \mid x - x_i \mid$.[11)] For simplicity we will assume that the distribution of preference along the unit line is uniform.

In the previous sections we have argued that the ability to attract broadband content for an ISP is a function of the number of connected customers. We have furthermore argued that the utility of a broadband ISP subscription is an increasing function of available broadband content. We will model this dependency in a very simplified way by assuming that when network i has n_i subscribers a certain amount of content is attracted and the utility of this on-net content is βn_i. As the number of subscribers changes, the value of available content changes according to the simple linear rule. The two competing broadband ISPs are however connected (either via the Internet as in figure 2, or via a direct interconnect link as in figure 3). A subscriber of ISP i will accordingly also get some benefit from the content available on the network of the competing ISP. When interconnect quality is k ($k \leq 1$) we will assume that the utility from the off-net content is $\beta k n_j$, where n_j is the number of subscribers on the competing network. The total content utility for a consumer connected to network i is accordingly assumed to be $\beta \cdot (n_i + k n_j)$, where $\beta > 0$ and $k \in [k,1]$.[12)]

Note that we add up the value of content on the two networks of our indirect network externality formulation. This formulation is only valid if there is no duplication of content. We are thus assuming that content providers at most place their content on one of the networks. This may be the case if it is costly to set up and run a broadband content service or if the broadband ISPs sign some exclusive deals with the content providers.

It follows from our formulation that when the quality of interconnect equals unity, customers are indifferent as to the distribution of off-net and on-net customers since on- and off-net traffic have identical quality. This is opposed to a situation where $k < 1$. Then, all other things being equal, a customer will prefer a network with many customers. When $k = \underline{k}$ the quality equals the quality available via the Internet (as in figure 2), whereas $k > \underline{k}$ implies that the two ISPs have agreed upon establishing an interconnect arrangement (as in figure 3) with superior quality. The parameter β is a measure of the de-

---

11) *This modelling is due to Hotelling. In the original Hotelling model, the horizontal differentiation is driven by transportation costs. The location of preferences is defined by the physical location of the customer in relation to the location of the suppliers.*

12) *We are here following Crèmer, Rey and Tirole (1999) by modelling network externalities such that consumers benefit from an increase in network size. They are also assuming that the additional utility stemming from an increase in on-net network members is larger than if the number of off-net networks members increases.*

pendency between the number of customers and the ability to attract interesting content to the local portal. When $\beta = 0$, available content and thus willingness to pay is independent of the number of customers. We are thus assuming that the two competing firms may differ in their ability to attract content, but this difference is entirely explained by differences in network size.

Finally we will assume that any consumer obtains net utility $v_i - p_i$ of being connected to broadband ISP i ($i = a,b$). The first term is a fixed advantage of being connected to broadband network i (which may differ between the networks). The second term is the monthly subscription fee paid to ISP i.

By collecting terms we obtain the following net utility function

$$U_i = v - t \mid x - x_i \mid + \beta \cdot (n_i + kn_j) - p_i$$

The customers' utility functions are accordingly assumed to be linear in consumption of the network service and money.

Assumption 1:
We assume that each of the customers along the interval [0,1] value the products sufficiently high such that they always prefer to subscribe to one or the other network. Thus, they will buy the subscription from either firm a or firm b, ie. the fixed advantage v of being connected to a network is sufficiently large.

Assumption 2:
In order to exclude the possibility of market cornering we assume that the magnitude of the indirect network externality β does not exceed the transportation cost t. This condition is fulfilled if $t > \beta\,(1 - \underline{k})$.

Assumption 2 is perhaps unfamiliar and may deserve a comment: Consider a symmetric case where the two suppliers charge the same price. Assume that almost all customers along the unit line, for some reason, are connected to supplier a. The marginal customer with the longest distance to travel to supplier a will compare the offer from the two suppliers and choose supplier a (and the market will accordingly be characterized by cornering) if: $\beta\,(1 + \underline{k}0) - t > \beta\,(0 + \underline{k})$ $\Leftrightarrow t < \beta\,(1 - \underline{k})$. Assumption 2 thus rules out the possibility of market cornering in such symmetric cases.

We define α as the market share of firm a. Assumptions 1 and 2 are then implying that $n_a = \alpha$, $n_b = 1 - \alpha$. For a given price vector, the location of preferences $x \in (0,1)$ for the consumer satisfying $U_a = U_b$ determines the market shares. The market shares of the two firms are accordingly:

$$\alpha = \frac{1}{2} - \frac{(p_a - p_b)}{2(t - \beta(1-k))}$$

$$1 - \alpha = \frac{1}{2} - \frac{(p_b - p_a)}{2(t - \beta(1-k))}$$

Note that if $k = 1$ and/or $\beta = 0$, the expression for market shares is identical to what we obtain in a standard Hotelling model (ie. a model without network externalities). By defining $\sigma = 1/2(t - \beta(1 - k))$ we can write the market shares:

$$\alpha = \frac{1}{2} - \sigma(p_a - p_b)$$

$$1 - \alpha = \frac{1}{2} - \sigma(p_b - p_a)$$

σ is a function of k where $\sigma(k) > 0$, $\sigma(1) = 1/2t$, $\sigma'(k) < 0$. Note that assumption 2 assures that $\sigma > 0$. The expression for market shares in our model is similar to the one we obtain in a standard Hotelling model with unit demand. In the standard Hotelling model, the parameter σ is interpreted as a measure of product substitutability. The products become closer substitutes if the transportation cost t between the two products is reduced. From our definition of σ it also follows that the products become closer substitutes in the eyes of the consumer if the quality of the link between the two networks is reduced. We can accordingly expect an increase in the cost of transport and an increase in the quality of the link between the two networks to have similar effects upon prices and profits.

A Two Stage Game

We are aiming at modelling a two-stage game. In the first stage the two ISPs set the interconnection quality k such that $\underline{k} \leq k \leq 1$. In the second stage, the two ISPs simultaneously set their prices for a given k.

Stage 2

In stage 2 of the game the quality of interconnect k is given and thus the measure of substitutability σ. The firms set their prices simultaneously, and firm i chooses p_i so as to maximize profits given by:

$$\pi_i = (p_i - c)\alpha_i = (p_i - c)\left(\frac{1}{2} - \sigma(p_i - p_j)\right)$$

where c is the unit cost. The first order condition for profit maximization is:

$$\frac{\partial \pi_i}{\partial p_i} = 0 = \frac{1}{2} - \sigma(p_i - p_j) - \sigma(p_i - c)$$

$$\Rightarrow p_i = \frac{1}{4\sigma} + \frac{p_j + c}{2}$$

Combining the first order conditions for firm *i* and *j* yields:

$$p_i = \frac{1}{2\sigma} + c \text{ and } \alpha_i = \frac{1}{2}$$

By inserting equilibrium prices and market shares in the profit function we obtain: $\pi_i(k) = (4\sigma)^{-1}$. From this profit function we can directly see that profit is a function of the level of product substitutability $\sigma = (2(t - \beta(1 - k)))^{-1}$ and thus of the interconnection quality. Inserting for σ and rearranging the profit function yields:

$$\pi_i(k) = \frac{1}{4\sigma} = \frac{(t - \beta(1 - k))}{2} \qquad (1.)$$

When $k = 1$, this profit function is identical to the one we obtain in a conventional Hotelling model with unit demand.

Stage 1

At stage 1 of the game the two firms decide whether to set up an interconnect arrangement or not. As already stated, stage 2 profit is a function of the quality of interconnection. Direct differentiation of the profit function (1.) with respect to *k* yields:

$$\frac{\partial \pi_i(k)}{\partial k} = \frac{1}{2}\beta \qquad (2.)$$

We readily see that in the symmetric case, the firms do not have conflicting interests with respect to network compatibility, ie. $\partial\pi_i / \partial k = \partial\pi_j / \partial k \ \forall k$. Thus, the two firms will always agree upon the optimal interconnection quality level *k*.

The effect upon profits from changing interconnect quality can be decomposed into a price and a market share (or volume) effect by differentiating: $\pi_i = \alpha_i (p_i - c)$:

$$\frac{\partial \pi}{\partial k} = \frac{\partial \alpha}{\partial k}(p_i - c) + \alpha_i \frac{\partial p_i}{\partial k}$$

The first term is the market share effect and the second term is the price effect. Consider first the price effect. By differentiating the equilibrium price we obtain:

$$\frac{\partial p_i}{\partial k} = \beta$$

The price effect of an increase in interconnect quality is accordingly always positive, and given that the ability to attract broadband content depends upon network size (ie. $\beta > 0$), the price effect is strictly positive. In the symmetric case considered in this section, market shares are unaffected by interconnect quality.

In the following we will apply the expression above to discuss the equilibrium quality of interconnect under various sets of assumptions.

Cost Free Interconnection Quality

We will first consider a situation where it is costless to improve the quality of interconnect.

From (2.) we have $\frac{\partial \pi_i}{\partial k} = \frac{1}{2}\beta \geq 0$ and thus the firms have no incentives to damage the quality of the link between the two networks. Furthermore, if possible, they have a mutual interest in improving the quality of this link. Then, both on-net and off-net traffic have the same quality level $k = \overline{k} = 1$.

Prices and profits increasing in the quality of the link between the two networks are due to two effects. The obvious effect is that for given market shares the willingness to pay increases for all customers as the quality is increased. Furthermore, when the quality of the link increases the competition between the two suppliers becomes less aggressive.[13] This is due to the fact that if the quality of the link is low, then a firm will get a large growth in profits by increasing its market share since it leads to a strong infra-marginal effect. The infra-marginal effect stems from the fact that the willingness to pay of existing customers increases more when market share increases when the quality of the link is poor, as compared to the situation where the quality of the link equals 1 (perfect interconnectivity). When the quality of the link equals 1, existing customers are indifferent with respect to market share. When comparing the conventional Hotelling model with our model featuring network externalities, the argument can be put the other way around: When the ability to attract content is a function of network size ($\beta > 0$) and networks offer less than perfect connectivity ($k < 1$), then the firms will compete more aggressively than what the conventional Hotelling model predicts.

Note that since the quality of interconnect in this case is costless the socially optimal quality evidently is $k = 1$. In the case without investment

---

13) *The best response functions ("reaction functions") in stage 2 of the game is: $p_i = R(p_j) = {}^1/_2(t - \beta(1 - k) + p_j + c)$. An increase in k will result in parallel shifts outwards for these best response functions and the firms do indeed become less aggressive as the quality of interconnect increases. We can furthermore see that R' = 0.5 – we are thus considering a stable Nash equilibrium.*

costs the market will accordingly implement the first best solution.

Convex Costs of Interconnection Quality

We will now consider a case where it is costly to set up the interconnection arrangement. The cost of interconnect in this case is the sum of the cost of investing in the necessary technical facilities and the transaction cost of writing a contract and the cost of mutual monitoring. It is reasonable that the total investment cost increases in interconnect quality. Furthermore, it may be reasonable to expect the cost to be convex, since, as interconnect quality increases, the complexity of the contract the two firms write becomes great. In particular the monitoring costs increase. As the quality of interconnect increases, the joint network of the two suppliers becomes more like a common facility where the firms have lots of opportunities for opportunistic behavior. When the quality of interconnect is high, a firm may for instance route traffic to the international Internet via the transportation network of the competing firm. In this way a firm can save infrastructure costs by congesting the competing network. Firms are of course not willing to agree upon interconnection unless the contract prohibits such opportunistic behavior. In order to observe and verify that the contract indeed is fulfilled, mutual monitoring is required.

Let the cost of investing in interconnect quality in order to increase the quality of interconnect k above $\underline{k}$ be $I = I(k)$, $I' > 0$, $I'' > 0$, $\lim_{k \to 1} I(k) = \infty$. Assume now that the two firms form an input joint venture where they equally share the cost of investing in interconnect quality. Each firm will then maximize the stage 2 profit minus the share of the interconnect cost the firm has to pay in stage 1. Thus the two firms will solve identical optimization problems and agree upon an interconnect quality level k^d characterized by:

$$k^d = \arg\max\left[\frac{t - \beta(1-k)}{2} - \frac{1}{2}I(k)\right]$$

Thus the joint venture investment level is characterized by:

$$\beta = I'(k)$$

We can compare this equilibrium quality level with the socially optimal quality. The first best interconnect quality, k^*, is defined as the quality level that maximizes customer gross surplus minus total production cost. Since the unit costs of serving customers in the two firms are identical and customers are distributed uniformly on the interval, first best is evidently characterized by sharing customers evenly among the two firms. Then average distance from the most preferred brand is 0.25. Inserting this yields:

$$k^* = \arg\max[v_0 - 0.25t + 0.5\beta \cdot (1 + k) - c - I(k)$$

The first best investment level is then characterized by:

$$0.5\beta = I'(k)$$

By comparing the first best investment level, we see that $k^* < k^d$ (since $\beta > 0.5\ \beta$). An input joint venture will thus choose a quality level of the interconnect arrangement exceeding the socially optimal level. The intuition behind this result is the following: There are two effects leading to the firms' stage 2 profits increasing in interconnect quality: The first effect is that for given market shares the willingness to pay increases for all customers as the quality increases. The second effect is that when the quality of the link increases, the competition between the two suppliers becomes less aggressive. Only the first effect is a social gain. Thus the input joint venture is over-investing in interconnect quality in order to reduce the stage 2 competitive pressure.

IV The Same Model Featuring Vertical Differentiation

The results obtained in the model above indicates that competing ISPs are always agreeing upon interconnect arrangements. In this section of the paper we will investigate whether this result stems from the strict symmetry in the model in the previous section. We will do this by introducing vertical differentiation into our model.

Vertical differentiation will arise when the two competing ISPs offering broadband access have asymmetric starting points with respect to the cost of upgrading customers (ie. one firm has to invest more than the other firm in order to upgrade the consumer with preferences located at the middle of the unit line). We will for simplicity capture this cost difference in the utility function by considering:

$$U_i = v_i - t\,|\,x - x_i\,| + \beta \cdot (n_i + kn_j) - p_i$$

The only modification compared to the utility function in the previous section is that the fixed advantage n_i, $i = a,b$, of being connected to network i is allowed to differ among the two suppliers. Define $\theta_i = v_i - v_j$. Given that the two firms charge identical prices the consumer with preferences located at the middle of the unit line will strictly prefer supplier i when $\theta_i > 0$. Thus the firm selling the superior product will obtain a market share above 0.5 when the two firms

charge the same price. In order to avoid corner solution we will have to add an additional assumption:

Assumption 3:
The valuation differential θ_i between products of the two firms is sufficiently low, such that there is one customer located at x, where $0 < x < 1$, who is indifferent to consuming the network service from the two firms, ie.: $| \theta_i | \leq 3\,(t - \beta\,(1 - k))$.

It follows from assumption 1 that the indifferent consumer located at $\alpha \in (0,1)$ derives a non-negative surplus. Given the assumptions above the market shares of firms a and b become:

$$\alpha = \frac{1}{2} + \frac{\theta_a}{2(t - \beta(1-k))} - \frac{(p_a - p_b)}{2(t - \beta(1-k))}$$

$$1 - \alpha = \frac{1}{2} - \frac{\theta_a}{2(t - \beta(1-k))} - \frac{(p_b - p_a)}{2(t - \beta(1-k))}$$

By defining σ in the same way as in the previous section the market shares can be written:

$$\alpha = \frac{1}{2} + \sigma\theta_a - \sigma(p_a - p_b)$$

$$1 - \alpha = \frac{1}{2} - \sigma\theta_a - \sigma(p_b - p_a)$$

The timing of the game we consider is identical to the game in the previous section. In the second stage, the two ISPs simultaneously set their prices for a given k and firm i chooses p_i so as to maximize profits given by:

$$\pi_i = (p_i - c)\alpha_i = (p_i - c)\left(\frac{1}{2} + \sigma\theta_i - \sigma(p_i - p_j)\right)$$

The first order condition for profit maximization is:

$$\frac{\partial \pi_i}{\partial p_i} = 0 = \frac{1}{2} - \sigma(p_i - p_j) + \theta_i\sigma - \sigma(p_i - c)$$

$$\Rightarrow p_i = \frac{1}{4\sigma} + \frac{p_j + \theta_i + c}{2}$$

Combining the first order conditions for firms i and j yields:

$$p_i = \frac{1}{2\sigma} + \frac{\theta_i}{3} + c \text{ and } \alpha_i = \frac{1}{2} + \frac{\sigma\theta_i}{3}$$

Assumption 3 assures that the indifferent customer indeed does exist, ie. $\alpha_i \in (0,1)$ implying the parameter restriction:

$$\alpha_i \in [0,1] \Rightarrow \left|\frac{\sigma\theta_i}{3}\right| \leq \frac{1}{2} \Leftrightarrow \left|\frac{\theta_i}{6(t - \beta(1-k))}\right| \leq \frac{1}{2}$$

From assumption 2 we have $t - \beta\,(1 - k) > 0$, and then the condition above is fulfilled when assumption 3 is fulfilled. By inserting equilibrium prices and market shares in the profit function we obtain:

$$\pi_i(\theta, k) = \frac{1}{4\sigma} + \frac{\theta_i}{3} + \frac{\sigma\theta_i^2}{9}$$

$$= \frac{(t - \beta(1-k))}{2} + \frac{\theta_i}{3} + \frac{\theta_i^2}{18(t - \beta(1-k))}$$

At stage 1 of the game the two firms decide whether to set up an interconnect arrangement or not. Direct differentiation of the profit function above with respect to k yields:

$$\frac{\partial \pi_i(\theta, k)}{\partial k} = \frac{1}{2}\beta - \frac{\beta\theta_i^2}{18(t - \beta(1-k))^2}$$

By definition we have $\theta_j = -\theta_j$, and thus we get:

$$\frac{\partial \pi_i}{\partial k} = \frac{\partial \pi_j}{\partial k} \forall k$$

This means that in a shared market equilibrium, the two firms will always agree upon the interconnection quality-level k even when the market share of the two firms differs.

By rearranging the differentiated profit function we obtain:

$$\frac{\partial \pi_i(\theta, k)}{\partial k} = \frac{1}{2}\beta\left(1 - \frac{\theta_i^2}{9(t - \beta(1-k))^2}\right)$$

The condition for having a shared market equilibrium is $| \theta_i | \leq 3\,(t - \beta\,(1 - k))$. This condition implies that the bracket above is positive. Thus in any shared market equilibrium profits of both firms increase in interconnect quality.

As in the previous section it can be useful to decompose the effect of changing interconnect quality by considering the price and the market share (or volume) effect. We recall that:

$$\frac{\partial \pi}{\partial k} = \frac{\partial \alpha}{\partial k}(p_i - c) + \alpha_i \frac{\partial p_i}{\partial k} \qquad (3.)$$

Consider first the price effect. By inserting the definition of σ in the equilibrium price and differentiating with respect to k we obtain:

$$\frac{\partial p_i}{\partial k} = \beta.$$

The price effect under vertical differentiation is positive and it is identical to the price effect in the symmetric case. This is opposed to the market share effect. Under vertical differentiation market shares become a function of interconnect

quality. By substituting for σ in the equilibrium market shares and differentiating we obtain:

$$\frac{\partial \alpha_i}{\partial k} = \frac{-\theta_i \beta}{6(t-\beta(1-k))^2} \qquad (4.)$$

The market share effect is positive for the firm selling the inferior service and thus it is negative for the firm selling the superior service. The negative market share effect for the firm selling the superior product is however dominated by the positive price effect as demonstrated above.

Although one of the firms has an advantage and a bigger market share, due to vertical differentiation, the incentives for interconnection quality do not differ between the firms. This result is in contrast to the results in eg. Katz and Shapiro (1985), Baake and Wichmann (1998) and Crèmer, Rey and Tirole (1999). In those papers, the dominating firm may have a lower incentive to increase compatibility (interconnection quality) than the smaller firm. Crèmer, Rey and Tirole (1999) assume that the firm preferring the lower k has the privilege to determine the quality of interconnection k.[14] This is in contrast to our model, where the difference in market shares stems from vertical differentiation. The difference in market shares in Crèmer, Rey and Tirole (1999) is due to differences in the exogenous given installed bases.

Cost Free Interconnection Quality

The differentiated profit function is everywhere increasing in k for both firms. Thus the firms have no incentives to damage the quality of the link between the two networks and furthermore, if possible, they have a mutual interest in improving the quality of this link. Then, both on-net and off-net traffic have the same quality level $k = \overline{k} = 1$. Thus when interconnect arrangements are cost free, a shared market equilibrium is characterized by the best possible interconnection quality. In the same way as in the previous section the chosen quality of off-net traffic is then identical to the quality of on-net traffic and the first best interconnect quality is chosen.

Convex Costs of Interconnection Quality

Consider now the case with convex interconnect costs, ie. we assume: $I = I(k)$, $I' > 0$, $I'' > 0$ and $\lim_{k \to 1} I(k) = \infty$. Since the firms do not have conflicting interests with respect to the optimal interconnect quality, it is reasonable to hold on to the assumption of an input joint venture where the cost of interconnect is shared equally among the two firms.

In market equilibrium the firms invest in an interconnect arrangement of quality:

$$k^d = \arg\max \left(\pi_i(\theta, k) - \frac{1}{2} I(k) \right)$$

The first order condition for optimal investments is then:

$$\frac{1}{2}\beta - \frac{\beta\theta_i^2}{18(t-\beta(1-k))^2} - \frac{1}{2} I'(k) = 0$$

$$\Leftrightarrow I'(k) = \beta - \frac{\beta\theta_i^2}{9(t-\beta(1-k))^2}$$

In Foros and Hansen (1999a) we show that this interconnect investment indeed exceeds the socially optimal investment for all $\beta > 0$.

V Conclusion

In this paper we have considered the incentives for Internet Service Providers (ISP) offering broadband access and portal services to strategically degrade the interconnection quality with the competitors. We have modeled this in a game where two firms choose the quality of interconnection before they compete over market shares à la Hotelling. In the case where there is no vertical differentiation, the firms split the market equally, and they have no incentives to degrade interconnection quality. Moreover, when interconnection is costly the firms will over-invest in interconnection quality as compared to the first best quality level.

We have also demonstrated that if the products from the two firms also are vertically differentiated, then the firm providing the superior product will have the larger market share. As far as interconnection quality is concerned, the results are however not changed. When the necessary conditions for a shared market equilibrium is fulfilled, the firms will agree upon an optimal interconnection quality. Furthermore, if interconnection quality is costly, the firms will agree upon a quality of interconnect exceeding the welfare maximizing quality level.

Initially we pointed at the choice of interconnection quality as an important strategic choice by

---

[14] *Assume the quality of the common interconnect arrangement is exceeding the quality level preferred by firm i. Then firm i has the opportunity of unilaterally degrading the quality of interconnect by manipulating the interface between network i and the interconnect facility. The firm preferring the lower interconnection quality can accordingly make a take it or leave it offer in negotiations over interconnection quality.*

ISPs offering broadband access and portal services. Our analysis indicates that if the ability to attract content is an increasing function of network size, and one expects market equilibrium to be characterized by more than one firm being active, then the strategy of degrading or choosing low interconnection quality is sub-optimal. In a duopoly, both firms will be better off by choosing high interconnection quality. With poor interconnection both firms will compete aggressively for market share because market share determines the value of the access product via available content. This is opposed to the case with high interconnection quality; then the two competing firms "share" content with each other since off-net content is available at high quality. Then market share is of less strategic importance resulting in an equilibrium where both firms compete less aggressively for market share resulting in increased profits for both firms.

References

Armstrong, M. 1998. Network interconnection in telecommunications. *The Economic Journal*, 108, 545–564.

Bailey, J, McKnight, L. 1997. Scalable Internet Interconnection Agreements and Integrated Services. In: *Coordinating the Internet*. Kahin, B, Kellser, J H (eds.). Cambridge, Mass., The MIT Press.

Baake, P, Wichmann, T. On the Economics of Internet Peering. *Netnomics*, 1, 89–105, 1998.

Cawley, R A. 1997. Interconnection, Pricing, and Settlements : Some Healthy Josting in the Growth of Internet. In: *Coordinating the Internet*. Kahin, B, Kellser, J H (eds.). Cambridge, Mass., The MIT Press.

Chinoy, B, Solo, T J. 1997. Internet Exchanges : Policy-Driven Evolution. In: *Coordinating the Internet*. Kahin, B, Kellser, J H (eds.). Cambridge, Mass., The MIT Press.

Choi, S Y, Stahl, D O, Whinston, A B. 1997. *The Economics of Electronic Commerce*. USA, Macmillan Technical Publishing.

Crémer, J, Rey, P, Tirole, J. 1999. *Connectivity in the Commercial Internet*. Toulouse, IDEI. (Technical Report.)

DangNguyen, G, Penard, T. 1999. *Interconnection between ISPs, Capacity Constraints and Vertical Differentiation*. Paper presented at the 2nd Berlin Internet Workshop, 28–29 May 1999.

Ergas, H. 2000. *Internet peering : A Case Study of the ACCC's Use of its Powers Under Part XIB of the Trade Practices Act, 1974*. Paper presented at the 3rd Berlin Internet Economics Workshop, 26–27 May 2000.

Foros, Ø, Hansen, B. 1999a. *Competition and Compatibility among Internet Service Providers*. Paper presented 26th annual E.A.R.I.E. Conference European Association for Research in Industrial Economics, 4–7 September 1999, Turin, Italy.

Foros, Ø, Hansen, B. 1999b. *Competing ISPs' incentive to increase interconnection quality*. Kjeller, Telenor Research & Development. (R&D Report R 22/99.)

Ims, L. 1999. Wireline broadband access networks. *Telektronikk*, 95 (2/3), 73–87.

Katz, M, Shapiro, C. 1985. Network Externalities, Competition, and Compatibility. *American Economic Review*, 75, 424–440.

Laffont, J, Rey, P, Tirole, J. 1998a. Network Competition I : Overview and Nondiscriminatory Pricing. *Rand Journal of Economics*, 29 (1), 1–37.

Laffont, J, Rey, P, Tirole, J. 1998b. Network Competition II : Price Discrimination. *Rand Journal of Economics*, 29 (1), 38–56.

Little, I et al. 2000. *Welfare Analysis of International and Domestic Markets for Internet Access Services*. Paper presented at the 3rd Berlin Internet Economics Workshop, 26–27 May 2000. (www.berlecon.de/ services/iew3)

Matutes, C, Regibeau, P. 1988. Mix and Match : Product Compatibility without Network Externalities. *Rand Journal of Economics*, 19, 221–234.

Matutes, C, Regibeau, P. 1996. A selective review of the economics of standardization – Entry deterrence, technological progress and international competition. *European Journal of Political Economy*, 12, 183–209.

Mueller, M, Hui, J Y, Cheng, C. 1997. The Hong Kong Exchange : The Economics, Evolution, and Connectivity of Asian Internet Infrastructure. In: *Coordinating the Internet*. Kahin, B, Kellser, J H (eds.). Cambridge, Mass., The MIT Press.

Rohlfs, J. 1974. A Theory of Independent Demand for Communications Service. *Bell Journal of Economics*, 5, 16–37.

Shapiro, C, Varian, H. 1998. *Information Rules : A Strategic Guide to the Network Economy*. Boston, Mass., Harvard Business School Press.

Market Managed Multiservice Internet

HUW OLIVER AND DAVE SONGHURST

Dr. Huw Edward Oliver (41) received his MA degree in Mathematics at Cambridge Uni. 1980, and his MSc (1985) and PhD (1988) in Computer Science at the Uni. College of Wales, Aberystwyth. He joined Hewlett-Packard Laboratories, Bristol in 1989 to work on Software Development Environments. Following a period at HP's Software Engineering Systems, Colorado in 1992 he returned to HP Labs in 1993 as Senior Member of Technical Staff and worked on real-time fault tolerant telecommunication systems. From 1996 to 1998 he worked on the ACTS ReTINA project; he has worked as Technical Reviewer of European ESPRIT and ACTS projects and is currently Manager of Hewlett-Packard's Internet Research Institute.

heo@hplb.hpl.hp.com

Dave Songhurst (52) is Research Consultant at BT Research. Until 1995 he led the Performance Engineering Section at BT Laboratories, working on traffic studies and network modelling and performance. In 1995 he left BT and worked for three years with Lyndewode Research, Cambridge, on the CA$hMAN project – a collaborative project in the European ACTS programme dealing with multiservice network charging schemes. He is now working with BT on the European M3I project – Market Managed Multiservice Internet. He is Editor of the book "Charging Communication Networks: from Theory to Practice".

dsonghurst@jungle.bt.co.uk

1 Introduction

Internet technology is becoming the infrastructure of the future for any information that can be transmitted digitally, including voice, audio, video and data services of all kinds. The traditional industries that supply such services are undergoing a major disruption. The new suppliers of such services are operating in a maelstrom. The problem has derived, in part, because of the historic roots of the Internet in the military and, later, academia. Neither of these communities had any motivation to build the infrastructure needed to price, charge and collect payments for the provision of Internet services.

Here is a simple overview of the problems some of these industries are facing.

1.1 Internet Service Providers

Internet Service Providers (ISPs) currently make money from

- Subscription charges (perhaps monthly for an individual; for a larger enterprise an annual fee plus initial connection charge plus customer premises equipment charge, etc.);
- Advertising;
- E-commerce (we are seeing the beginnings of this);
- Technical Support.

In some countries (for example, the UK) ISPs receive a 'kickback', that is a portion of the charge for telephone access from the telecommunication operator. The telcos often view this as an abuse of data service lines (set up for telephone banking, for example) but have not received a very sympathetic ear from the regulators who want to see Internet usage increased.

But the ISPs only rarely make their money from the actual services they provide. Currently ISPs provide access and some basic services but those are becoming commoditised. Above the network service, ISPs would like to provide advanced services (VoIP, Intranet, Extranet, integrated e-mail/voice messaging, content, relationship management, etc.) and would like to build sophisticated business models around those services.

Another disruption will occur when network quality of service mechanisms start to appear. Then there will be many customers with business critical needs who will willingly pay for the better quality of network service. ISPs are, in general, unprepared for this event.

Finally, the current peering models of carrying peer ISP's traffic free are breaking down. The asymmetric volumes of traffic are leading larger ISPs to refuse to carry the traffic from smaller ones for free.

1.2 Application Server and Service Providers

The computing industry is being transformed by the ability to outsource computing services over the Internet. Five years ago running a large piece of business software meant making large investments in installation, maintenance, upgrading and deployment and support staff. Now services are being provided on a pay-per-usage basis. Application Service Providers (ASPs) own the application (complex data analysis tools or sophisticated business software) and provide it on-line to a wide range of different enterprises. Sometimes a separate ASP will host the application and provide network access to it.

Their problem is that charging for these services has to be done in an ad-hoc manner. Quality of service is being provided, but in a way that is exclusive to the customers of the ASPs.

The ASPs would like to provide their customers with flexible ways to pay for the services. At the moment the model is similar to traditional software licensing – pay a fixed fee whatever your level of use. More desirable is usage charging both for datasets and for level of speciality of software. This would allow smaller suppliers to enter the market more easily and would allow customers to tailor their use to their budget.

1.3 Telecommunication Operators

For the telecommunication operators the move to an IP based infrastructure is an opportunity as much as a problem. Current billing systems are

- Centralised (data is transported from around the network to a central processing farm via physical tape or X.25 or FTAM);

- Inefficient (bill production takes months);
- Expensive.

The PSTN billing systems have typically taken hundreds of man-years to construct. Re-implementing that software for the Internet would be a nightmare. For the telcos, the Internet philosophy of intelligence at the edges applied to billing is intriguing.

Finally, PSTN Telephony Traffic Theory has been good for half a century. The single service connection-based telephone networks and basic human behaviour have not changed significantly during that period. This all changes with the multiple voice/video/data services of the packet switched Internet (and will get worse when multiple qualities of service arrive).

1.4 Mobile Telephony Operators

Enormous investments are being made in the future of the packet switched mobile services (GPRS and UMTS). In May 2000 the UK auctions for UMTS bandwidth raised a total of 22.48 billion pounds. How will such investments be recouped?

Current charging models for mobile communications are inappropriate for mobile packet networks. The reason is that these are connection-oriented and only take into account the duration of the connection through the access network. Such schemes do not account for the actual traffic that goes through the connection and work as if the connection was always fully utilised. The wrong incentives that these pricing schemes present to the users are to keep connections only during the time data are being transmitted, and hence incur a high signalling overhead for setting and tearing down connections at the time scales of the bursts of the data.

While traditional connection charges may be used in the initial absence of other mechanisms, this position is unsustainable.

The problem, common to all these industries, is that an inappropriate pricing system will convey the wrong incentives to the end users and lead to inefficiency, reduced profitability and ultimately lead to congestion.

2 Solutions through Network Charging

The introduction of network charging aims to tackle these problems. It has (at least) three inter-related objectives:

- to differentiate services, giving users the option to pay more for specific services or for better service quality;
- to ensure network efficiency through suitable incentives so that users will constrain their demands appropriately;
- to ensure fairness, so that network resources are shared according to need.

There are three types of charging mechanism that could be used, possibly in combination, in order to compose a sound market-driven resource allocation mechanism: differentiated service charging, usage-based charging, dynamic charging.

2.1 Differentiated Service Charging

Customers are able to subscribe to different services at differentiated prices depending on the facilities being offered or the level of service quality. Pricing may be flat rate or combined with usage charges. Differentiated flat rate pricing is a way to achieve market segmentation but still suffers from the problems with regard to efficiency and incentives. It may be difficult to offer high quality services economically unless some form of usage charging is used to constrain demand.

2.2 Usage-based Charging

Telephony networks have traditionally used duration-based charging with distance-related tariffs. The development of multiservice packet-switched networks based on asynchronous transfer mode (ATM) led to proposals for a range of new services, including variable bit-rate services with varying levels of service guarantees. Research into new charging schemes for variable bit-rate traffic led to the European ACTS project CA$hMAN, which implemented and trialled a range of novel charging schemes (Songhurst, 1999). These included tariffs based on a combination of duration and volume charges for each connection, designed to approximate the “effective bandwidth” of the connection. Effective bandwidth (Kelly, 1997) is a measure of the resource that the network needs to reserve for a variable bit-rate connection in order to ensure the required service quality.

In the Internet usage-based charges have taken the form of charges for access time and charges for total data volume. However, unlike ATM the Internet protocol is not connection-oriented and so does not facilitate the concept of reserving resources for individual connections This makes it difficult to relate usage charges to service quality for individual connections. Developments such as resource reservation protocol (RSVP) are an attempt to counter this, but there are other reasons for seeking a different approach in the Internet:

- The Internet is inherently decentralised – the concept is a simple network with intelligence existing in the end-systems.

- There is an endless possible variety of user applications that will use the Internet. It is not feasible to second-guess these and offer differentiated network services that will match their requirements.

- Bandwidth is cheap. Intense competition has driven ISPs to flat-rate, or free, Internet service, at least for low-rate access.

These factors lead towards a decentralised and dynamic approach to charging and service quality.

2.3 Dynamic Charging

The marginal cost of carrying data over the Internet when there is capacity available is essentially zero. However with no usage charging there is no way to control congestion and ensure that users can get good quality service. This leads to the concept of charging for usage only when there is contention for resources. The charge that users should pay is a *shadow price* that reflects the cost to other users through increased congestion.

This principle has been recognised in telephone networks through the simplistic approach of peak and off-peak pricing. However, considerations of competition and network efficiency should lead to a genuinely dynamic charging system where charges for usage vary in real time in relation to congestion in the network. The Internet does offer the possibility of a simple means to achieve this through the use of packet marking (Gibbens and Kelly, 1999).

3 M3I – Market-Managed Multiservice Internet

In January 2000, Hewlett-Packard Ltd, Athens University of Economics and Business, Technical University of Darmstadt, BT, ETH Zurich and Telenor began a two year project called M3I to build and test network service charging approaches.

3.1 Objectives

In the M3I project we propose the use of pricing mechanisms for controlling demand for scarce resources, in order to improve the economic efficiency of the system. Standard results in economic theory suggest that increasing the value of the network services to the users is beneficial to both the users and the network operator (since he can charge them more and get back a bigger percentage of their surplus). Using pricing mechanisms helps in that respect. When demand is high, prices are being raised and hence deter the users with low valuation for the service to use it. This leaves resources to be available for the users that value them more, and hence are ready to pay more.

As noted earlier, prices can work in many time scales. *Dynamic prices* work in short time scales and reflect the instantaneous state of the network in terms of excess demand for resources. These are in many cases called *congestion prices* since they reflect the cost due to performance degradation that a user imposes to the other users that share the same network resources. The work in Gibbens et al. (1999) is an attempt to define congestion prices in an implementable way for packet networks (the Internet), where these prices achieve the economically fair allocation of the resources of the network to the competing connections, see Kelly et al. (1998). Prices that work in larger time scales (time-of-day pricing) can also be used in our framework. The idea is that differential pricing makes users "self-select" and choose the service quality they prefer for the posted price.

In general, the difficulty of designing an appropriate pricing and charging system comes from the need to solve three problems simultaneously. The solution must make a good economic solution, technical solution and end-user solution. Taking these separately:

3.1.1 Economic Issues

The solution must provide the correct incentives, be fair and be the basis of a sustainable business. It must also be sufficiently flexible to allow new business models to be created around it.

We have seen the high cost of putting the current telecommunications billing systems in place. We must be sure to avoid doing this again. The issue is particularly acute for any measurement-based approaches. While Internet measurement techniques have advanced in recent years, capturing Internet traffic behaviour analytically is notoriously difficult. The speed, computational and storage requirements of measuring traffic at a low level and recreating high level statistics are largely unknown.

If dynamic pricing is used there needs to be a clear and auditable relationship between the advertised current price of service use and the final customer bill.

When end-systems choose their own rate control algorithms to suit a wide variety of different applications, will the network operate in a stable way? We need a solution whereby rates converge to stable values rather than fluctuate randomly or cyclically.

3.1.2 Technical Issues

Internet congestion effects take place at many different levels of timescale. Many periods of congestion have a duration of just a few milliseconds. These are far too short to be dealt with by any changes in users' behaviour. There is also the issue of duration of user flows. These are often of too short a duration to include any user reaction to network state. Studies have been done, however, to show that the presence of such short flows do not lead to instability (Kelly 1999).

The primary thrust of a market managed approach is to use price as an admission control to the network. It needs to be provably effective if it is to replace the traditional approaches. These two forms of admission control will also live side by side for the foreseeable future. What will be the global effects of having both price and technical admission control for different flows?

Flow measurement indications and price signals will lead to extra control traffic. How significant will this be? We must ensure that we have a scalable wide area measurement solution.

Finally, we have to be clear about what congestion means. Flows typically cross multiple hops across multiple domains. They are not connection oriented and different packets within a flow may take different routes. The term "network congestion" is an oversimplification when the reality is that we have a widely distributed set of transiently congested resources.

3.1.3 User Issues

Perhaps the most widely levelled criticism of usage charging and dynamic pricing is that the user cannot predict the final bill. Internet usage is increasing dramatically and hence users worry that a usage based charge might lead to a similar dramatic increase in their bills. Studies have shown that predictability has a high value (UC Berkeley 2000).

A second issue is that of relating user level service utility, such as sending an email or attending a teleconference, to lower level service charging. A packet charge, megabyte charge or even flow charge means little to the average end user. The end user needs to be able to see the overall price attached to their high level service in order to decide whether to use it.

Pricing and charging schemes aim to shape the end user's behaviour to one that leads to high network efficiency and maximum social utility. There is, however, often too great a disconnect between the user's control and how an application generates traffic. We need a solution that puts bandwidth consumption control in the user's hands simply and effectively.

Finally, we need a solution that will work for the many different kinds of user. The main categories are that of residential user and corporate user. While a residential user might be very sensitive to charges, a corporate user will only be indirectly motivated to limit costs. We need a solution that allows a corporate policy to be enforced without overly restricting the individual corporate users.

3.2 A mechanism for Congestion Pricing

Ramakrishman and Floyd (1999) have proposed the implementation of Explicit Congestion Notification (ECN) in TCP. Resources that are nearing congestion can signal this fact to end-systems by marking packets before it is necessary to start dropping packets. The authors of this proposal suggest that end-systems should back-off in response to marked packets. However Gibbens and Kelly (1999) propose that congestion marks should be interpreted as charges, and end-systems should be free to vary their rate in accordance with their utility and the charges received. Various algorithms for marking packets are possible – a simple method is to mark all packets arriving at a resource when its input buffer exceeds a given threshold.

With congestion charging, end-systems will no longer use the standard TCP rate control algorithm. Instead they will use algorithms that depend on the needs of the user application with parameters that could be set by an end-user. One can distinguish at least three broad types of application having different rate control requirements:

3.2.1 Adaptive Applications

Many applications are able to adapt their rates over a broad range without a major loss of utility to the end-user. For such adaptive applications Kelly et al (1998) propose a variant of TCP rate control in which one parameter can be set by the application, in response to network charging signals, so as to ensure convergence to a rate that maximises the user's net utility. This 'willingness-to-pay' parameter identifies the end-system's desired level of payment per unit time, and available bandwidth is shared between flows in proportion to these parameter values.

3.2.2 Real-time Non-adaptive Applications

Applications of this type, such as high-quality audio and video, are unable to adapt below a certain rate. They will exercise rate control to maintain at least this minimum rate regardless of pricing signals unless the charge becomes so large that the flow decides to terminate. An application initiating a new real-time flow will exercise a form of self-admission control by testing the

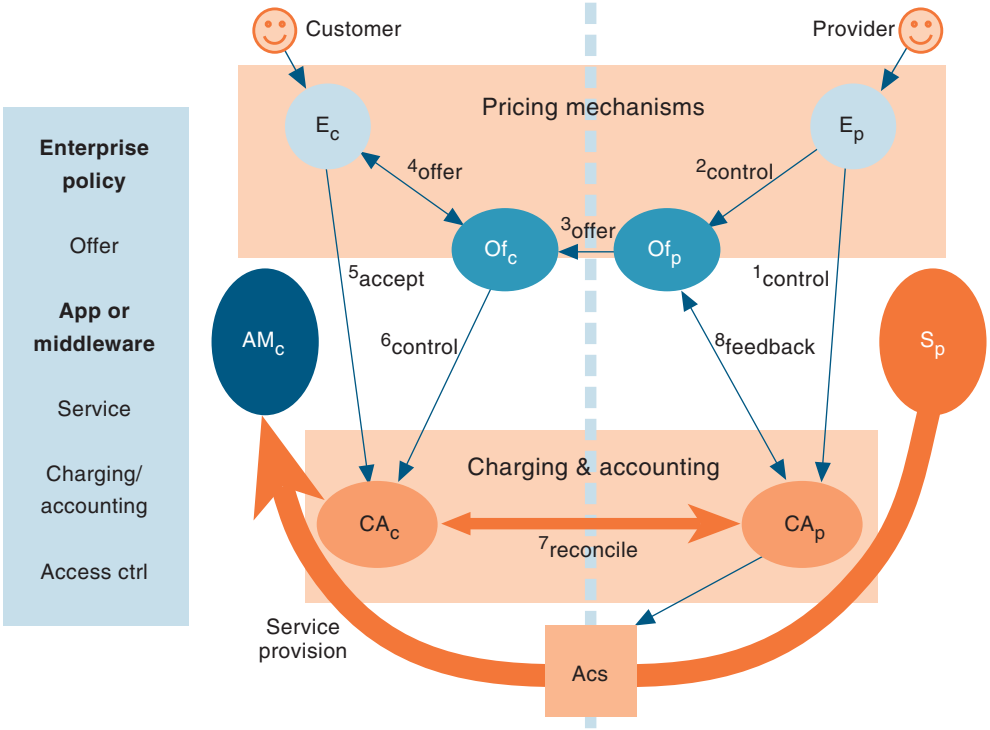


Figure 1 Architecture for market managing Internet services

network bandwidth price before deciding whether to commence. Gibbens and Kelly (1999b) also discuss the use of gateways between a group of users and the network, able to use congestion pricing information to perform a distributed acceptance control function.

3.2.3 File Transfers

File transfers require to transfer a file of a given size within as short a period as possible, generally with no benefit to the end-user before it is completed. Examples of this type of connection include traditional file transfers (transfers of data files using ftp applications), email, and many components of Web pages (image files, etc). File transfers do not benefit from shared bandwidth – their optimum rate control is to transmit at peak rate when the price is low enough, otherwise to suspend transmission. Existing TCP rate control does completely the wrong thing for this wide class of application. Gibbens and Kelly (1999) and Key and Massoulié (1999) discuss possible file transfer algorithms with congestion charging.

The key issue is that congestion charging enables rate control, and hence network resource allocation, to be determined intelligently by end-systems according to their application requirements and utilities.

3.3 Architecture

The architecture consists of the following main pieces:

- Pricing and Admission Control. Here are the mechanisms for setting prices for services, communicating those prices to the end user and the APIs that allow the user to react to that pricing information.

- Charging and Accounting. Here are the mechanisms to measure service usage and to charge according to the prices communicated as above. The aim will be to ensure that the accounting system is as generic as possible, with the ability to plug in appropriate meters that are specialised to measure different aspects of the network and of services provided over the network.

- Below these the Internet is providing raw services. The mechanisms for providing those services at different qualities are still under development in the IETF and one of the challenges for the architecture is to be neutral toward ultimate choices of QoS mechanism as far as possible.

Another challenge is that of scalability. We propose that edge providers charge their local customers for both sending and receiving (more accurately for each class of service in each direction at a separate price). Thus, we extend charging recursively to apply at the boundary between any pair of providers. We also claim that as long as all customers are usage-charged for both sending and receiving by the network providers at all edges, both multicast and unicast charging can be achieved very scalably. While Figure 1

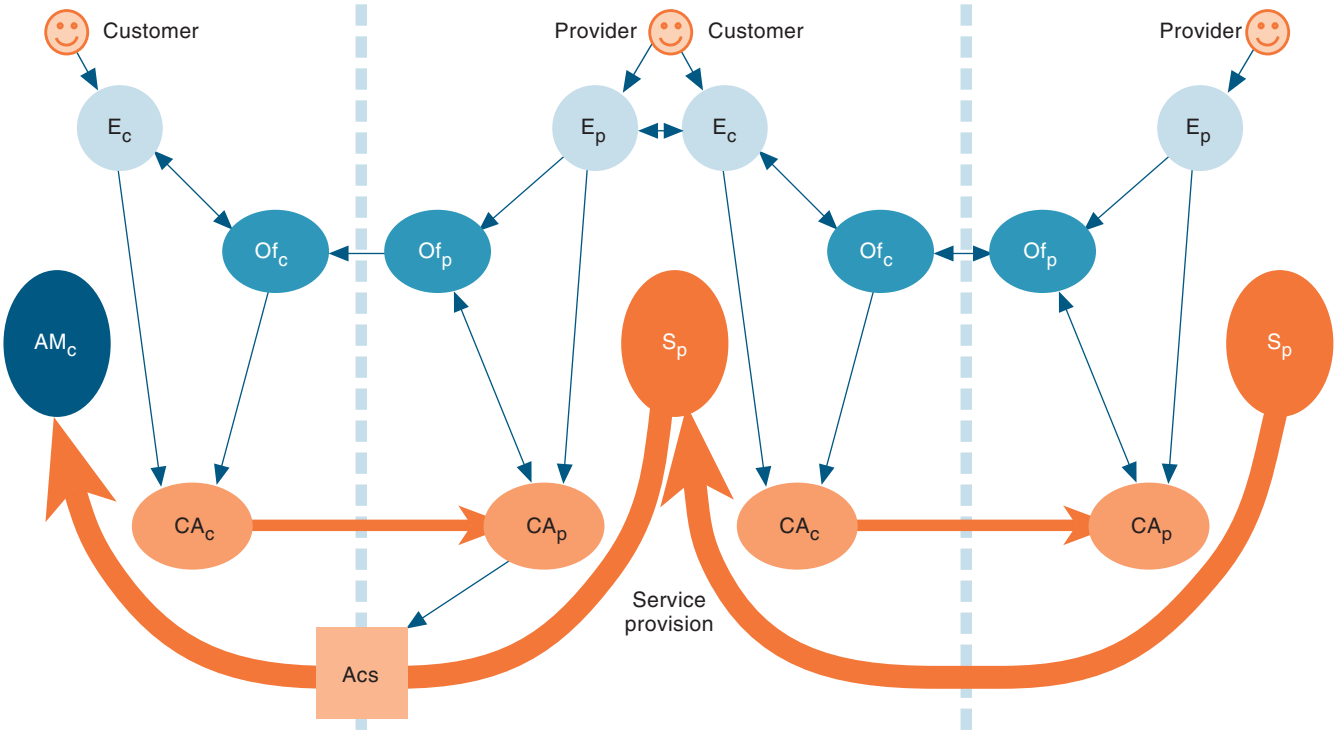


Figure 2 Inter-provider relationship as a recursive version of the architecture

represents the customer-provider relationship at the network edge, it can also represent the relationship between network providers. In this case the 'customer application', AMc, would be just another network service in a chain ending eventually at an edge-customer, but otherwise the architecture is recursive as shown in Figure 2.

3.4 Planned Approach

M3I is tackling the following tasks in order to pursue the objectives described above.

3.4.1 Test Platforms

The project will develop test platforms that will be used to evaluate the above architecture, including the implementation of dynamic charging schemes.

3.4.2 Trials

The test platforms will be used to carry out trials, including user trials, which will investigate issues such as

- User sensitivity to quality and price, and user/network interface required to communicate this;
- Use of agents to automate user reaction to dynamic pricing;
- How corporate customers can manage price reaction policies for groups of users.

3.4.3 Modelling

Modelling is a major part of the project, supplemented where possible by results from trials. It includes the following:

- Development of cost models for Internet service providers;
- Formulating business models for ISPs who offer differentiated services through charging;
- Analysis and simulation of networks using dynamic charging in order to evaluate stability;
- Economic models of competitive markets involving many ISPs having differing business models and using different charging schemes. An important aim is to analyse the competitive advantages of a dynamic market-based approach to charging.

4 Summary

There are currently many proposals for pricing and charging models but, in the absence of real experience of providing commercial communication services for a wide variety of applications, the proposals are generally based on plausibility arguments. Within M3I we intend to carry out large scale simulations and experiments and modelling analyses to gather information about the applicability of differentiated pricing through usage-based and dynamic charging schemes.

Our key goal is to gauge the effectiveness of market forces in allocating network resources to end-systems according to their application requirements and utilities.

Partners

The M3I project is a joint undertaking by key members in the Internet Industry: Hewlett-Packard Ltd, Athens University of Economics and Business, Technical University of Darmstadt, BT, ETH Zurich and Telenor.

References

Gibbens, R J, Kelly, F P. Resource pricing and the evolution of congestion control. *Automatica*, 35, 1969–1985, 1999.

Kelly, F P, Maulloo, A, Tan, D. Rate control for communication networks : shadow prices, proportional fairness and stability. *Journal of the Operational Research Society*, 49, 237–252, 1998.

Courcoubetis, C A, Reiman, M I. Pricing in a large single-link loss system. In: *Proc. 16th International Teletraffic Congress.* Elsevier, 1999, 737–746.

MacKie-Mason, J K, Varian, H. Pricing the Internet. In: *Public Access to the Internet.* Kahn, B and Keller, J (eds.). Prentice Hall, 1994.

Odlyzko, A. *A modest proposal for preventing Internet congestion*. (2000, September 04) [online] – URL: http://www.research.att.com/~amo/ doc/modest.proposal.ps.

Songhurst, D J (ed.). *Charging Communication Networks : from Theory to Practice.* Amsterdam, Elsevier, 1999 . (ISBN 0-444-50275-0.)

Kelly, F P. Charging and accounting for bursty connections. In: *Internet Economics.* McKnight, L W and Bailey, J P (eds.). Cambridge, MA, MIT Press, 1997, 253–278.

Ramakrishnan, K, Floyd, S. *A proposal to add Explicit Congestion Notification (ECN) to IP.* The Internet Society, January 1999. (RFC2481.)

Gibbens, R J, Kelly, F P. Distributed connection acceptance control for a connectionless network. In: *Proc. 16th International Teletraffic Congress.* Elsevier, 1999, 941–952.

Key, P B, Massoulié, L. *User policies in a network implementing congestion pricing.* Workshop on Internet Service Quality and Economics, MIT, December 1999. (2000, September 04) [online] – URL: http://research.microsoft.com/research/network/ publications/ISQElm.ps.

(2000, June 29) [online] – URL: www.index.berkeley.edu

Kelly, F P. *Models for a self-managed Internet.* Royal Society discussion meeting on Network Modelling in the 21st Century, London, December 1999. (2000, June 29) [online] – URL: http://www.statslab.cam.ac.uk/~richard/research/topics/royalsoc1999/

The Internet Market Structure: Implications for National and International Regulation*)

ØYSTEIN FOROS AND HANS JARLE KIND

Øystein Foros (31) holds a Master of Business and Economics degree (1993) from the Norwegian School of Management (BI) and is Cand. oecon. (1996) from the Norwegian School of Economics and Business Administration (NHH). He has been employed as Research Scientist at Telenor R&D since 1996. Since 1998 he also holds a position as Scientific Advisor to the Foundation for Research in Economics and Business Administration (SNF). Current areas of specialization are industrial organization analysis of the Internet and telecommunication markets.

oystein.foros@telenor.com

Hans Jarle Kind (41) holds a Master of Business and Economics degree (1993) from the Norwegian School of Economics and Business Administration (NHH). His principal employer is the Foundation for Research in Economics and Business Administration (SNF). Presently he holds a post doc. position in economics at NHH/Norwegian Research Centre in Organization and Management (LOS), and his current areas of specialization are international trade, industrial organization and competition policy.

Hans.Kind@nhh.no

1 Introduction

Internet connectivity may be seen as a composite good that is produced by the complementary inputs local and global access. In addition to the infrastructure components, software and content components affect the customers' demand for connection to the Internet. The value for consumers lies in the whole chain or system, and not in a particular segment. E-commerce without network access, for instance, is valueless. Similarly, the majority of the new multimedia services have limited value for the users without a broadband local access network.

In this paper we will concentrate on the infrastructure components local and global access sold to residential consumers as an example of the interplay between firms in complementary sub-segments, and how it affects the optimal regulation policy. In such context, a low regulated price of access to domestic bottlenecks is good news for firms in complementary segments, but we show that it may be detrimental to national welfare.

In the first part of this paper we provide a description of the market structure for Internet connectivity outside the US. In the unregulated retail market Internet Service Providers (ISPs) offer Internet connectivity to the end-users. The ISPs purchase the complementary essential inputs local and global access. The telecommunication incumbents seem to have market power in the input segment for local access, while the global infrastructure is controlled by a limited number of Internet Backbone Providers (IBPs).[1)] Regarding regulation policy, the prevailing paradigm in the EU is to constrain the incumbents controlling the local loop to use cost-oriented access prices.

In the second part we argue that this policy may be inferior for a government in an open economy that seeks to maximize national welfare. If a cost-oriented access price should be optimal domestic policy, it would have to be true that foreign firms have no market power in essential complementary segments. For electronic communication services this assumption will rarely hold. If foreign input providers in complementary segments have market power, the domestic authority should set the price on local access above costs.

A cost oriented regulation may *de facto* commit the domestic authority to set a low fee on local access, leading foreign firms to charge prices above those that would be the outcome in an unregulated market economy. Thereby regulation may result in an excessive profit flow to the foreign country. This underlines the importance of taking into account how foreign firms react when the competition policy is designed, and further makes it clear that there may be a need for an international competition policy. In the input market for local access both the price and quality dimensions are regulated in most countries. The other input, global access, is currently unregulated. However, a two-sided price regulation that sets the price of both local access and global access may be harmful to national welfare if the global access provider vertically integrates into the domestic downstream market. The reason is that if the access price is set too low, the vertically integrated global access provider may engage in non-price discrimination of the national firm. Somewhat surprisingly, this need not harm the domestic incumbent, but rather lead to even higher consumer prices than we would observe in an unregulated market economy.

Unregulated retail market for Internet connectivity is a key assumption in our analysis. This assumption corresponds to what is commonly observed.

The article is organized as follows. In Section 2 we consider some key features of the market for communication services in general, and the market for Internet connectivity in particular. In Section 3 we use a simple model to highlight how the special Internet market structure may affect public regulation policy. Section 4 concludes.

*) *This research has been partly financed by The Research Council of Norway. Research Grant No. 137521/510 is gratefully acknowledged. The authors thank Bjørn Hansen for helpful comments.*

1) *Also in complementary segments such as content and software, there seems to be large companies with considerable market power.*

2 Market Structure

In this section we will briefly discuss some key features of the market for Internet connectivity sold to end-users outside the US. The purpose is to make a baseline for our analysis in the next section.

2.1 Current Regulation in Telecommunications

There are large fixed costs and natural barriers to entry in major segments of the telecommunication industry, and in a free market economy we would expect a few firms to have considerable market power. It is therefore not surprising that firms controlling essential bottlenecks in this industry have been subject to comprehensive public regulation.

The retail market for Internet connectivity is currently unregulated in most countries, while the input segment for local access is regulated both with respect of price and quality. According to Laffont and Tirole (2000) the regulators' decision not to regulate the retail market builds on two premises. First, if the local bottleneck is eliminated, then head to head competition in the retail market ensures that there is no need for regulation. Consequently, regulation of the local access bottleneck is sufficient to ensure competition in the retail market. Second, the products and services in the retail market change very fast, and this makes it very costly to monitor the retail markets compared to the wholesale markets.

The former argument deserves a comment, since local access is not the sole input bottleneck for retail ISPs producing Internet connectivity. Consequently, there is an asymmetric regulation regime in two dimensions. First, the fact that the end-user market is unregulated creates an incentive for a vertically integrated provider of local access to discriminate against rivals in the retail segments. This issue has been examined by several researchers, see eg. Laffont and Tirole (2000). Second, since the market for global access is unregulated, there is also an asymmetry between the two complementary infrastructure inputs, and this will be of particular interest due to the US dominance in provision of the unregulated input (global access).[2)] We focus on the interplay between the regulated local input and the unregulated global input.

The prevailing regulation regime of local access in Europe is cost-oriented, which means that the incumbent is not allowed to charge higher access prices than those reflecting its long-run marginal costs.[3)] The incumbent controlling the local telephone network often uses three main arguments against cost-based regulation. The first argument is that it is practically impossible to compute the long run marginal cost in an industry involving large joint costs. The second argument is that the local access network for telephony no longer constitutes a bottleneck, because cable-TV and wireless networks are bypass opportunities for residential users.[4)] The third argument is that a cost oriented regulation will reduce the incumbent's dynamic incentives to invest into infrastructure and product innovation. The danger that regulation creates dynamic inefficiency is an important topic in all technologically advanced industries. These issues are discussed in detail by Laffont and Tirole (2000). We will not go into this discussion, which basically has the same arguments whether we consider the traditional market for telephony services or the market for Internet services.

In addition to the analysis of the interplay between the providers of the complementary inputs, we focus on the timing of the interaction between the domestic regulator and the market players. The current sector specific cost-based price regulation for local access is often seen as a "hands-on" *ex ante* approach, while the competition rules are seen as an *ex post* regulation approach. This distinction may be misleading, since the current cost-based sector regulation *de facto* often will appear as an *ex post* regulation (see eg. Laffont and Tirole (2000) and Foros, Kind and Sørgard (2000)). Our purpose is to compare a situation where the domestic regulator credible commits to a given price policy for local access before the input suppliers choose their wholesale prices with a situation where the domestic regulator does *not* make such a commitment. The former we refer to as *ex ante* regulation, while the latter we refer to as *ex post* regulation. Hence, in our context *ex ante* regulation does not imply cost-based price of local access.

---

2) Except for local access, all the major bottlenecks seem to be controlled by American firms. Cisco Systems, for instance, is the main supplier of the routers in the basic infrastructure, MCI WorldCom is the dominating provider of global access to the Internet backbone, Microsoft is close to a monopolist in providing PC operating systems, and AOL/Time Warner will presumably control an important part of the content market.

3) See Mognes and Nord (1999) for an overview of current and future regulation of local access regulation with focus on the Norwegian market.

4) There may be a contradiction in these two arguments since large common costs obviously may give rise to bottlenecks.

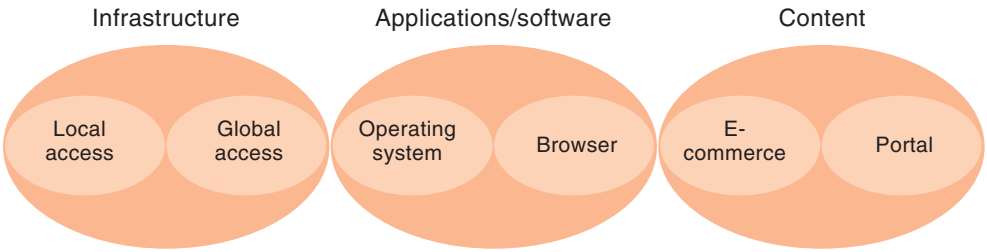


Figure 1 Structure of the Internet market

2.2 Complementarity between Segments

Figure 1, where we have distinguished between infrastructure, applications/software and content, illustrates the chain of complementary inputs. The important point here is that consumer demand is for the whole chain or system, not for a single component, and that the profit level that a firm can extract depends both on the competition in its own segment and the competition in complementary segments. Hence, a firm can extract a larger share of the total value when the complementary segments are highly competitive or strongly regulated. Note, however, that the value of such a chain can be extracted only once.

The complementarity between the Internet segments is important even for firms that have limited market power. For example, the lower the price of local access, the higher the price that Amazon.com can charge their consumers. A strict price regulation of local access is thus good news for Amazon.com, and this is true irrespective of the competitive pressure from other booksellers operating through the Internet. Moreover, a firm such as Microsoft, that has monopolized one component (the operating system), has no incentive to integrate into complementary segments where we are likely to have perfect competition or strictly regulated prices. Microsoft may, on the other hand, have incentives to enter unregulated market segments with imperfect competition.[5)] In this sense it is not surprising that Microsoft has avoided the transport segment, but become an important player in the market for Internet browsers.[6)]

It is hardly the welfare of consumers Bill Gates (Microsoft) and Steve Case (AOL) have in mind when they argue that the prices charged for telephone lines should be reduced (Davos meeting, February 2000). Rather, it is the fact that the telephone networks are complements to a large share of the goods sold by Microsoft and AOL. These companies will thus extract a larger share of the total Internet industry profit if telephone prices are reduced.

2.3 The Retail Market for Internet Connectivity

Figure 2a shows a situation where Internet connectivity and local access is sold bundled to the end-users by local Internet Service Providers (ISPs), who in turn buy local access as an input from a Local Access Provider (LAP) and global access as an input from an Internet Backbone Provider (IBP). In that case regulatory pressure has obligated the LAP to open its network and sell high-quality local access as an input to potential rivals in the end-user market at a price equal to long run marginal costs.[7)] In Figure 2b, in contrast, the LAPs sell local access as a separate service to end-users, who possibly buy Internet connectivity from another firm. The end-user then combines the two components into the composite good. These are presently the two dominating market constellations, and the discussion in this section holds independent of which of these constellations we consider.[8)]

---

5) *For a further discussion, see Economides (1998c).*

6) *An open question is to what extent we will see vertical integration between infrastructure providers of Internet connectivity and content and applications providers (see Figure 1). The experience from the past indicates that there will be few such mergers. Current high-level services are rarely bundled and sold together with connectivity from ISPs (Clark, 1999a). Recently we have however, seen several proposals for mergers between facility-based firms and content providers.*

7) *It should be noted that cable-TV companies do not face this regulatory requirement, and that is presumably one reason why they have so far not offered local access to independent ISPs.*

8) *In Section 3, where we discuss optimal regulation policy, we will focus on the situation described in Figure 2a.*

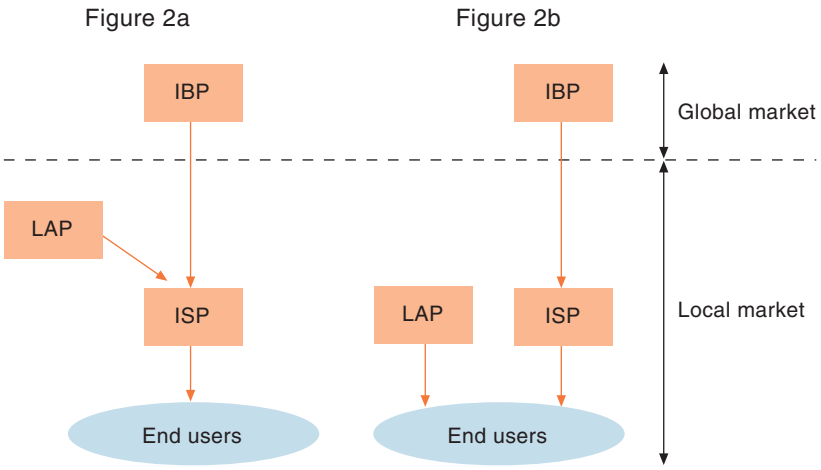


Figure 2 The major market constellations for Internet connectivity

2.4 Local Access

Local access, or the "last mile into homes", seems to be a bottleneck component for all electronic communication services sold to the residential market. Obviously, and in contrast to other potential bottlenecks, local access has to be offered locally. Presently, there are at most two incumbents that can provide wireline local access in a given area; cable-TV companies (using cable modems over hybrid fiber coax) and telephone companies (using dial-up modem, ISDN or ADSL through the copper pair into homes). The high up-front investments of new wireline facilities, and the possibility of increasing the capacity and quality of existing local telephony and cable-TV networks, indicate that there will be no rush to enter this market and install additional wires to homes (Clark, 1999a). The competitive pressure in this segment is thus limited. Furthermore, cable-TV providers have not been offering local access as an input to independent ISPs. Therefore, the telephony incumbent has been the only provider of local access as an input to independent ISPs.

Technical innovations may certainly alter this picture, but the costs of installing additional wires to the homes are likely to be high also in the foreseeable future. Hence, some wireless technology with enough bandwidth seems to be one of the more promising technological innovations that may alter the picture. Several serious technical challenges exist (such as the scarcity of spectrum resources) that must be solved before companies offering wireless local access will be able to capture significant market shares. Note, however, that the technical problems regarding LMDS now seems to be solved, and broadband access is currently offered via this technology.[9] Another alternative may be some kind of combination of different technologies, or what is often called hybrid technologies. One possibility is to use satellite in one direction and telephone copper lines in the other direction.[10] Independent of the technological evolution, it is realistic to believe that the incumbent who is controlling the copper pair into homes will still have considerable market power in the input market for local access. The reason is that the incumbent has access to every potential consumer, since almost all private homes have a telephony subscription from this incumbent.

2.5 Global Access

Global access to the core Internet Backbone is controlled by a few firms, called Internet Backbone Providers (IBPs), that sit on the top of the hierarchy. In addition to giving access to information located on servers in the US, the input from the IBPs also secure access to the core routing structure, and access to all Internet addresses in the world (Milgrom et al., 1999 and Srinagesh, 1997). A limited number of core IBPs co-operate in creating a consistent routing structure. These arrangements are called peering, and a key feature of these arrangements is to create full routing tables. The full routing tables is a part of the input supplied to regional ISPs, and they define the addresses that can be reached (Srinagesh, 1997).

The US government was the main financial sponsor of the Internet when the system was in its infancy. At that time the suppliers of Internet services seemed to be more interested in building new infrastructure and attracting new consumers on a non-profit basis, and the norm was cost-free interconnection between all ISPs. As the Internet matured the US government's position as financial sponsor ended. In 1993 the US government decided to leave the management of the core Internet Backbone to commercial IBPs, and this has resulted in a more commercial and hierarchical structure. While the core IBPs still have cost-free interconnection between each other, they now charge smaller regional ISPs for access to their global infrastructure and core routing services. In other words, the smaller regional ISPs have become customers (or resellers) of the core IBPs facilities and services. An example of this is UUNET (an MCI WorldCom subsidiary), who ended the cost-free interconnection regime in 1997 and started to charge smaller ISPs for access to their backbone.[11]

---

[9] *Local multipoint distribution systems (LMDS) is a fixed wireless technology.*

[10] *For a further discussion of these issues, see Clark (1999a, 1999b) and Speta (2000).*

[11] *See Mackie-Mason and Varian (1997), Bailey and McKnight (1997), and Werbach (1997) for a summery of the Internet's history.*

The core global infrastructure is controlled by a small number of American IBPs, such as MCI WorldCom, Sprint, GTE, AT&T and Cable & Wireless, that control between 85 % and 95 % of the total backbone traffic in the US (see Cremer et al. (1999) for an overview). It should be noted that global access is much more essential for Internet connectivity than for conventional telephone services: while only a relatively small portion of world wide telephone calls go to the US, the majority of the Internet traffic has to go through the US. For the location of Internet facilities we thus have a clear asymmetry between the US and the rest of the world.

> "In the good old days, network engineers didn't connect with another company; they connected with another engineer whom they knew and trusted. These "peering arrangements" typically were informal agreements to exchange traffic without money changing hands. But as the industry matures, settlement-free interconnect does not necessarily provide appropriate incentives to the industry players. "Why should I help my competitors by giving them free access to my network?" say the suits. "But the Internet won't work unless everything is connected to everything else," say the geeks. Both are right. Interconnection is healthy for the industry as a whole, but the current business model for interconnect may easily generate incentives for individual carriers to overcharge their competitors." Hal Varian (1998).

The core American IBPs are subsidiaries of the major facility-based telecommunications firms, and they also control the majority of the transatlantic lines. When European, Asian or Australian ISPs wish to connect to one of the US backbone, they must usually pay for communications both ways. Even if each IBP separately has limited market power, a group of co-operating IBPs may have considerable market power (Cremer et al., 1999, Milgrom et al., 1999). For instance, although the quality of the backbone is enhanced through the IBPs efforts to coordinate their core routers, it is clearly tempting to use this formalized co-operation as a collusive device (Varian, 1999). Recently, we have also observed that IBPs have vertically integrated into the retail market for Internet connectivity (the ISP segment) also in Europe. It is well known that when an input-segment monopolist integrates into a competitive retail-segment, it may have incentives to practice foreclosure against rivaling firms in order to give its own subsidiary a competitive advantage.[12)]

2.6 Vertical Integration

Above we noted that both LAPs and IBPs recently have integrated vertically into the ISP segment. Even though both the LAPs and the IBPs control essential inputs, they do not have the same possibilities to engage in foreclosure. The reason is that the LAPs typically are subject to regulation of price and quality for local access as an input component, while there is no regulation of the global access input controlled by the IBPs. The domestic incumbent that is controlling the local bottleneck comes from the regulated telecommunication world, while the IBPs who are controlling the global bottleneck come from the unregulated Internet world. The ability and incentive for a dominating IBP to practice foreclosure was given attention during the MCI-WorldCom merger proceeding. It was pointed out that foreclosure could take a number of forms, including price increases on global access for the ISPs and non-price discrimination through quality degradation of the inputs. The EU is now raising the same concerns regarding the proposed merger between MCI WorldCom and Sprint.[13)]

For the EU it is not sufficient that the firms controlling a global input component are complying with US antitrust laws, and this fact has received some attention lately. This has led to discussions of whether there is a need for some kind of global regulation, which for instance may take place

---

12) *Another reason for the IBPs to vertically integrate into the retail market is the trend towards relatively more local Internet traffic. This tendency is probably due to new consumer-types and new services in the Internet. In non-English speaking countries, content intended for the mass-market must be produced locally or translated. Furthermore, for new interactive services, such as telemedicine, tele-education, and video conferencing, a larger portion of the communication is probably between consumers in the same geographical area than what is the case for conventional Internet services such as web browsing. Hence, it may be important for the IBPs to be active in the local market since the importance of local traffic as a strategic variable increases (see Foros and Hansen, 1999a, 1999b).*

13) *After this article was written the merger was dropped due to resistance from the US government.*

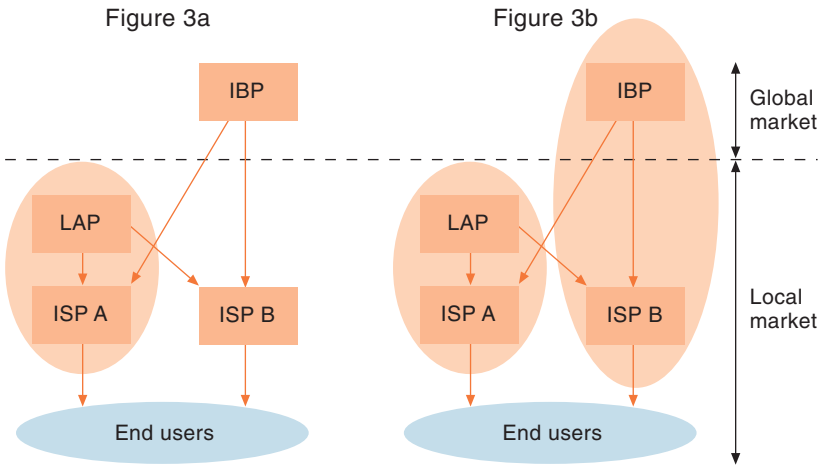


Figure 3 Vertical separation versus vertical integration

through WTO agreements that reduce the scope for foreign firms to utilize their global market power excessively.

Most Internet users have previously demanded services such as e-mail and web-browsing. These services are not particularly sensitive to delays. The quality requirements have changed, however, not least because there is an increasing demand for interactive video that requires real time communication. The market thus becomes more sensitive to the quality of interconnection, and, hence, quality degradation can be an important strategic weapon. Integrated firms such as MCI WorldCom may therefore gain an advantage by offering premium services to the consumers of its own subsidiaries, and this fact has received a great deal of attention.[14] In such circumstances it is almost impossible for the regulation council to decide whether the integrated firm offers new functionality (higher quality) based on technological advantages, or whether the firm practices non-price discrimination through quality degradation of the input sold to the rivals in the retail ISP-segment. The Microsoft case gives an illustration of the problem in such contexts (see Economides, 1998b). As long as the IBPs has incentives to practice foreclosure, it is hard to limit their ability to do so.

The telecommunication industry has a long history of using foreclosure as a strategic weapon. The Bell System (later AT&T) provides a good example of this. Some one hundred years ago the Bell System was the dominating telephone company in the US. Though it faced competition from a large number of local operators, it was the technical leader for long-distance calls. Bell refused to give the local rivals access to the national long-distance network, and could thereby offer a superior service to their own end-users. As a consequence, the local competitors were soon driven out of the market.

3 Regulation of Local and Global Access. Who gains?[15]

Few sectors have historically been so intensively regulated as the telecommunication industry, and the market structure we have described above indicates that the need to regulate will not be lower as the Internet becomes increasingly important. Below we will present a stylized model where we explicitly take into account that a few American firms dominate some of the market segments that are complementary to local access, and show how this may affect the optimal regulation policy outside the US. We will also discuss the scope for international regulation policies.

The discussion in this section is purely verbal, and is partly based on a formal analysis in Foros, Kind and Sørgard (2000). The analysis is made as simple as possible in order to highlight how the Internet market structure affects the regulation policy. We have thus made no attempt at being "realistic", but rather set up a framework that allows us to point out some potentially important aspects of the interplay between the regulator and the firms providing local and global access.

Figure 3 provides a simple illustration of the vertical market structures for Internet connectivity that we will consider. In Figure 3a the LAP is vertically integrated with ISP A, while ISP B is an independent national firm. These firms in turn buy access to the global backbone from the IBP. In Figure 3b both the LAP and the IBP are vertically integrated, with ISP A and ISP B, respectively.

3.1 One-sided Cost Oriented Regulation of Local Access

In this section we discuss the effect of a one-sided regulation of the local access price. In a context like Figure 3a, where the IBP operates as wholesaler of global access, a regulation that sets the price for local access equal marginal

[14] *See Shapiro and Varian (1998) for a discussion of the MCI WorldCom case, and Economides (1998a, 1998b) and Foros, Kind and Sørgard (2000) for formal analysis.*

[15] *This part is based on Foros and Kind (2000).*

costs is only optimal if the IBP has no market power. If the IBP vertically integrates into the retail market, as illustrated in Figure 3b, a cost-oriented regulation will never be optimal.

3.1.1 The Upstream Backbone Provider is Non-integrated

In this subsection we will assume that the LAP is vertically integrated with ISP A, while ISP B is an independent firm. Domestic shareholders own these firms, while foreigners own the IBP. We will first consider the case where the ISPs can costlessly connect to the backbone, and this is consistent with the "old" regime in the Internet. Recently, however, we have observed that the IBPs have begun to charge the ISPs for connectivity to the backbone, and presumably this pricing behaviour will become more widespread along with the increased commercialization of the Internet (Frieden, 1999). We will therefore also consider the case where the IBP chooses the price of backbone connection so as to maximize profit. It will be shown that the optimal regulation policy depends crucially on whether the backbone connection price is endogenously determined or not.

Exogenous input price for global access

Suppose that the LAP has a long-run marginal cost equal to c_l, and that it operates ISP A as a subsidiary. The LAP may also sell local access to ISP B, in which case ISP A and ISP B compete in the end-user market. The end-user market in turn consists of a number of consumers that differ in their willingness to pay for connection to the Internet. With a large number of consumers we will then have a downward sloping demand curve like *DD* in Figure 4, which measures quantity on the horizontal axis and price on the vertical axis. For the moment we will assume that access to the global backbone is costless. What is the optimal price strategy of the LAP?

Access to the local loop is an essential input, and the LAP is therefore able to foreclose ISP B. This can easily be done by setting a price of local access so high that ISP B cannot operate profitably.[16] In that case the integrated LAP avoids price competition in the end-user market, and will be in a position to charge the monopoly price p_m and sell Internet access to q_m consumers. Thereby the LAP maximizes its profit level, which can be expressed as $\pi_{LAP} = (p_m - c_l)q_m$. This in turn is equal to total domestic profits π_D (since ISP B is inactive), and is illustrated by the blue quadrant in Figure 4.

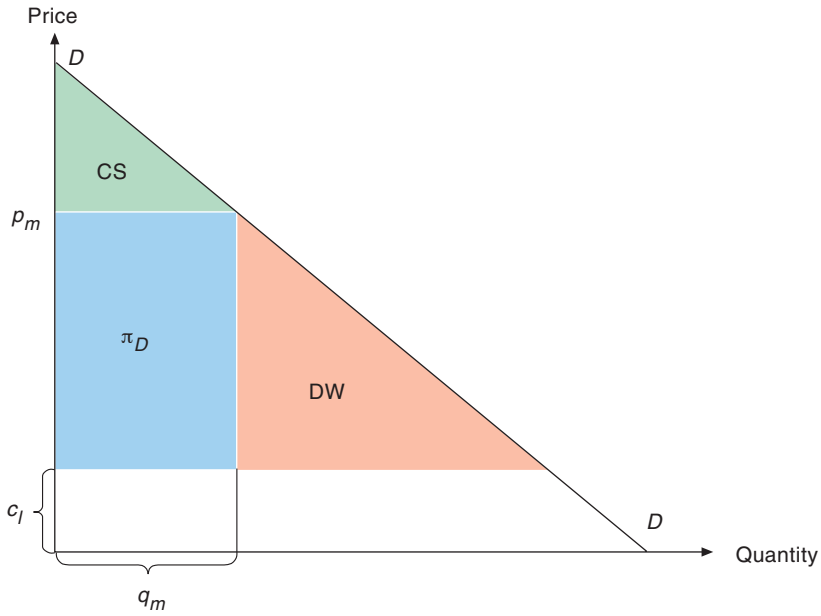


Figure 4 Lack of competition creates a deadweight loss

Consumer surplus (*CS*) is equal to the green area in Figure 4. Since the domestic firms are owned by national shareholders, we will follow the standard procedure in economics and measure welfare as the sum of consumer surplus and domestic profits, $W = CS + \pi_D$ (the green and blue areas in Figure 4). Consequently, we do not value consumers higher than shareholders, or vice versa. This is not a critical assumption; the qualitative results go through even if we use different weights.

It is well known that monopoly pricing is socially inefficient, and the size of the inefficiency – the deadweight loss – is equal to the red tripod DW in Figure 4. The inefficiency arises because the LAP has to consider two effects of a change in the consumer price. Suppose, namely, that the LAP charges a price p' and sells Internet access to q' consumers. By charging a somewhat lower price the LAP will sell Internet access to one more consumer, and in isolation this increases the profit level by $(p' - c_l)$ units. In that case, however, it will also have to reduce the price on all the other q' units that it sells. These effects cancel each other at the monopoly price, while the latter (former) effect dominates for lower (higher) price levels.[17] In a socially efficient solution, however, the LAP should serve all consumers who are willing to pay at least c_l.[18]

---

[16] *In order to make the discussion as simple as possible, we assume that the ISPs offer homogenous goods in this section. Product differentiation will be discussed in Section 3.2.*

[17] *Implicitly we have here assumed that the monopolist is unable to price discriminate.*

[18] *The profit loss that arises if consumers are charged less than p_m is just a transfer of money from domestic firms to domestic consumers. This is irrelevant from a social point of view, since CS and π_D are given equal weights in W.*

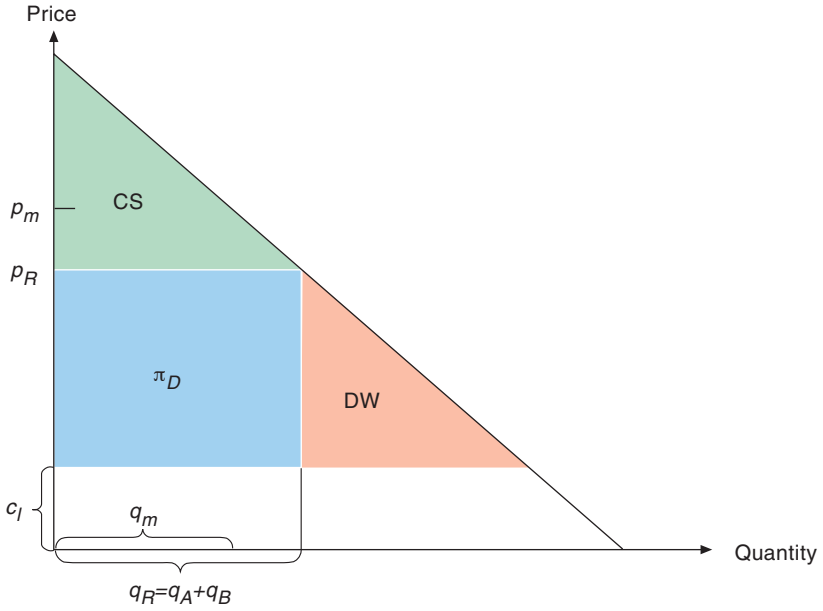


Figure 5 The deadweight loss is reduced due to higher competition

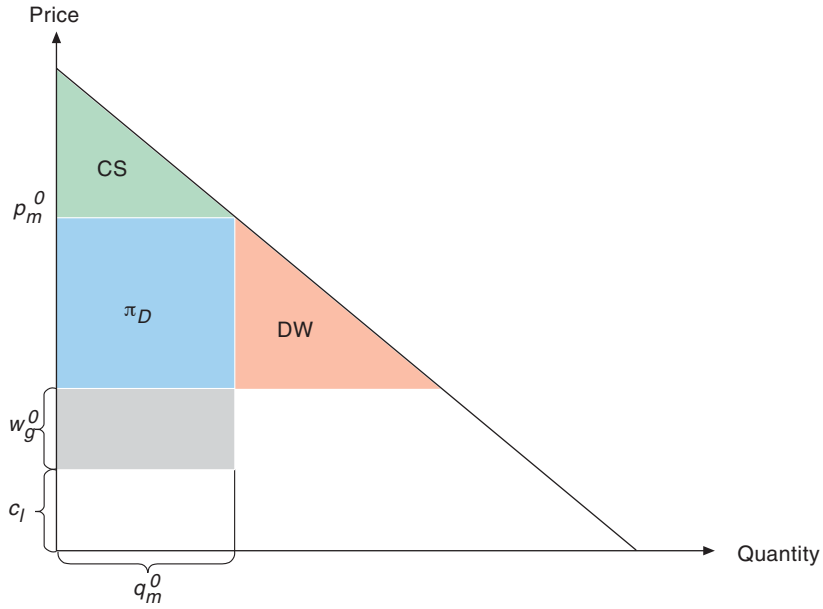


Figure 6 Market equilibrium with endogenous price from the IBP

There will be no deadweight losses if there is perfect competition, but this is obviously not true for the telecommunication market. In an ideal world the government could replicate a perfectly competitive market by requiring that the consumer price for Internet connection is set to c_l, but that policy is not used in practice (see Laffont and Tirole, 2000). The prevailing regulation regime is instead to regulate the price that the bottleneck owner charges from other firms in the industry. The regulated price will typically be lower than the price that forecloses other firms in the industry, but never lower than long-run marginal costs. In the present context the former implies that both ISP A and ISP B will sell Internet connection, while the latter implies that the integrated LAP earns a non-negative profit on sales to ISP B. The consumer price will subsequently be determined by competition between ISP A and ISP B in the end-user market.

The less expensive it is for ISP B to connect to the local loop, the lower the price it charges the consumers. Which price, then, should the regulator set for access to the local loop (given that it cannot be lower than c_l)? The answer to that question is simple; the price should be set equal to c_l. The reason is that the regulator aims to minimize the deadweight loss, and thus the consumer price should be as close to c_l as possible.[19)]

Figure 5 illustrates the new situation, where we have assumed that the consumer price is reduced from p_m in the unregulated economy to p_R when the price of access to the local loop is regulated. With this lower price more consumers are willing to buy access to the Internet, and the demand for Internet connection has thus increased from q_m to q_R. The total profit level is reduced, so that the blue area π_D in Figure 5 is smaller than in Figure 4 (it is impossible to generate a higher profit level than the one a monopolist obtains). Note also that π_D now is split between the integrated LAP and ISP B, so the profit level of the integrated LAP is significantly reduced due to the regulation. However, the lower consumer price implies that the deadweight loss is reduced. Therefore regulation increases domestic welfare, and it follows that the higher consumer surplus more than outweighs the loss of domestic profit.

Endogenous Price of Global Access

Above we argued that the integrated LAP has incentives to utilize its market power in the local loop and foreclose the independent Internet Service Provider, ISP B. However, we also have a second essential input, namely access to the global network. This fact was not visible in the former analysis, since we assumed that the ISPs had free access to the backbone. Free access used to be the norm, but casual observations show that this is no longer the case

[19)] *Note that even if ISP B pays only c_l for access to the local loop, the consumer price will be higher. The firms will therefore make some profit, and this is likely to be true unless there is a very large number of competitors.*

(Frieden (1999), and Cremer et al (1999)). We will therefore extend the analysis, and assume that the IBP determine the price for backbone access so as to maximize its profit level.

It is obvious that the end-user price increases when the IBP uses a positive mark-up on its long-run marginal costs. How large the mark-up will be depends on how price sensitive consumer demand is, but the basic analysis in the unregulated market economy will be exactly the same as the one we used above. The integrated LAP will still have incentives to foreclose ISP B in order to avoid competition, and there will be a deadweight loss that calls for regulation of the price for local access. The outcome in the unregulated economy is illustrated in Figure 6, where p^0_m is the monopoly price that the LAP charges in the end-user market, q^0_m is the corresponding consumer demand, and w^0_g is the price that the IBP sets for access to the global backbone. Total revenue to the IBP is equal to w^0_g times the number of consumers connected to the Internet, q^0_m. This is represented by the grey square in Figure 6. Since the IBP is a foreign firm, its revenue does not count in the measure of domestic welfare. We thus still have that $W = \pi_D + CS$, ie. the green and blue areas.

The domestic deadweight loss occurs for exactly the same reason as in the first analysis; the integrated LAP charges an end-user price (p^0_m) that is higher than its long-run marginal costs ($c_l + w^0_g$). What is interesting is how the regulation policy should be designed when w_g is endogenous. We will consider two cases. First we will consider so-called *ex post* regulation. This implies that the regulator first observes the market prices for local and global access. If the regulator finds that the price of local access is not optimal, it can intervene and change that price. Secondly, we will consider *ex ante* regulation, whereby the regulator commits itself to using a certain price. Incumbent telecommunication companies repeatedly complain over the *ex ante* regulation, but below we will argue that there are strong arguments in favour of this regulation regime.[20)]

Ex Post Regulation

With *ex post* regulation – which is the dominating competition policy in most markets in both the EU and the USA – the regulator first observes the market equilibrium, and then decides whether it should intervene. Foreign firms lie outside the regulator's jurisdiction, and in the present context the regulator therefore has the same problem as in the earlier analysis: it has to determine the socially optimal access price to the local loop (ie. the price that the integrated LAP is allowed to charge ISP B). In the following it is convenient to denote this access price by w_l.

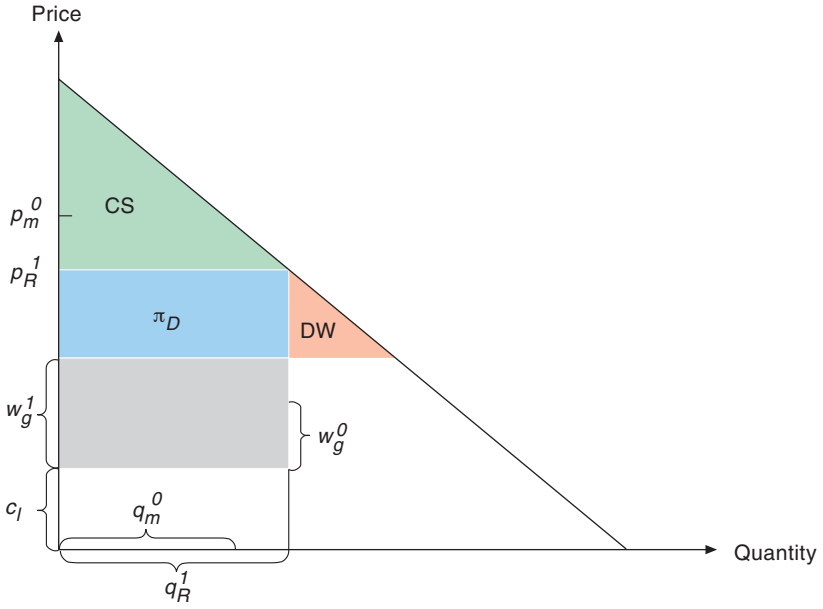


Figure 7 Ex post regulation may reduce domestic welfare

It is actually very simple for the regulator to find the optimal value of w_l. Whatever price the IBP charges, the best the regulator can do is to set $w_l = c_l$. If it sets a higher value of w_l, the consumer price – and thus the deadweight loss – will be unnecessarily large. Figure 7, where p^1_R and q^1_R denote the end-user price and quantity, illustrates the outcome. As expected, the deadweight loss and the consumer price are reduced ($p^1_R < p^0_m$), since regulation ensures competition between ISP A and ISP B. However, note also that the access price to the global backbone is now higher than in the unregulated economy; $w^1_g > w^0_g$. The reason is that the IBP is aware of the fact that the regulator *ex post* always prefers $w_l = c_l$. This means that the consumer prices will be relatively low in any case, and therefore the IBP has an incentive to increase the price it charges from the integrated LAP and ISP B (and capture some of their profits). We may well expect that this effect more than outweighs the higher consumer surplus, in which case domestic welfare ($W = \pi_D + CS$) falls as a consequence of the regulation policy. The net result of regulation is thus to increase the profit flow to foreign country excessively, and the domestic country would be better off in the imperfect and unregulated market economy. This result is, to the best

20) *Whether we have* ex ante *or* ex post *regulation does not matter if the price of access to the global backbone is exogenous. The reason for this will become clear later.*

of our knowledge, new to the literature. However, it reflects a basic insight that has been stressed in the political economic literature: what matters is not whether the market economy is imperfect, but whether the government can do better.

Ex Ante Regulation

From the above it is clear that *ex post* regulation may be inferior, because part of the initial domestic profit is shifted to the foreign IBP without ensuring a sufficient increase in consumer surplus. The prevailing regime in the EU, however, is in principle to use so-called *ex ante* regulation for the telecommunication industry. In this sense there is an asymmetry, because most other industries are subject to *ex post* regulation.[21)]

With *ex ante* regulation the regulator commits itself to set a certain price for access to the local loop. Which price should the regulator choose? Obviously not $w_l = c_l$; we have already seen that that is detrimental to welfare, and any value of w_l below the one chosen by an unregulated LAP increases profit shifting to the foreign firm. In fact, we cannot preclude the possibility that the best the regulator can do is to accept relatively high consumer prices, and allow the integrated LAP to become a monopolist in the end-user market. In that case the outcome is the same as we had in Figure 7, where ISP B is foreclosed and the consumer price equals p^0_m. Note, however, that the regulator must *credibly commit* itself not to intervene. The reason is that the regulator *ex post* always will have an incentive to set $w_l = c_l$. Since the IBP is well aware of this fact, a simple announcement not to intervene is implausible. Whether the EU regulation policy is credible in this sense is open to debate, but for the rest of this paper we will take the intentions of the EU seriously, and only consider *ex ante* regulation.

3.1.2 The Upstream Backbone Provider is Vertically Integrated into the Retail Market

The result that the integrated LAP forecloses ISP B illustrates a quite general insight from economic theory, namely that parent companies have incentives to give their subsidiaries a competitive advantage. This may be a particularly fruitful strategy to the extent that the parent controls an essential input, as is the case for local access providers. However, also the backbone providers control an essential input, and recently we have seen that the core American IBPs have vertically integrated into the market for Internet connectivity also in Europe. This raises two important questions that we will highlight in this section. First, how is domestic welfare affected if a foreign IBP vertically integrates into the ISP market? Second, what are the implications of this for the optimal regulation policy?

In order to answer these questions, we will assume that the IBP is vertically integrated with ISP B as illustrated in Figure 3b. This means, *de facto*, that the integrated LAP now faces competition from an Internet Service Provider that has cheap access to the backbone (at a cost equal to the IBP's long-run marginal costs, rather than w_g). The competitive pressure will therefore be higher than when the IBP was not integrated, and this in turn implies that consumer prices will be lower.[22)] Consumer surplus will therefore increase. Indeed, we may also expect domestic welfare to increase, even though the profit of ISP B now accrues to the foreign firm (so that welfare is now equal to $W = CS + \pi_{LAP}$). The reason is that the IBP and the LAP in the present context are in a symmetric position, controlling each their essential input and each their end-user provider. Therefore we may foresee more aggressive competition between these firms, leading to a significant reduction of the welfare loss compared to the situation where only the LAP was vertically integrated.

What is more surprising, perhaps, is how the regulator will set the access price to local loop in this case.[23)] In Section 3.1.1, where only the LAP was vertically integrated and ISP B was an independent national firm, we argued that the LAP's choice of access price to the local loop might be optimal also from a social point of view, even though it meant that ISP B would be foreclosed. The rationale for this "non-intervention" policy was to avoid excessive profit flow to the foreign firm. In the present case foreigners also own ISP B, and a low access price to the local loop will therefore directly transfer profit to foreigners (this is true even if the price from the IBP were exogenous). Nonetheless, the regu-

---

21) *The reason why the EU uses* ex ante *regulation is quite different from the one that we will stress below, and some pros and cons of that policy are discussed by Laffont and Tirole (2000).*

22) *The fact that vertical integration may reduce consumer prices is well known in the economic literature, and goes under the name "avoidance of double marginalisation". See Spengler (1950) and Economides and Salop (1992).*

23) *We assume* ex ante *regulation, and thus implicitly that the regulator is able to credibly commit itself. Otherwise we now from the previous discussion that the regulator ends up with a regulated local access price equal to long-run marginal costs.*

lator may now prefer a lower access price than the LAP. The reason is that the IBP's response to a lower local access price is distinctly different when it is vertically integrated. First, the access price w_g is now less important, since it does not affect the costs of ISP B. Second, as argued above, the competition will be more aggressive when both the bottleneck owners are vertically integrated. This is beneficial for the consumers, and makes the regulator more willing to set a low access price to the local loop.

To sum up, we thus argue that it may be good news for the home country if foreign IBPs vertically integrate with domestic ISPs, and that the socially optimal access price to the local loop may then be lower than if the ISPs are independent.

3.2 Two-sided Global Regulation

Even though it may be favourable with a relatively low access price to the local loop when the IBP is vertically integrated, it will still be optimal for the regulator to set an access price above long-run marginal costs to avoid excessive profit shifting to the foreign country. This raises the question of whether there is a need for a global price regulation, which for instance may take place through WTO agreements that reduce the scope for foreign firms to abuse their international market power. In order to discuss the welfare effects of international price regulations in a meaningful way, however, we must take into consideration the fact that firms have other strategic choice variables than price alone. For the telecommunication industry it seems particularly relevant to consider whether there are incentives for using non-price discrimination of interconnecting competing firms. More specifically, in this section we will discuss whether a price cap on backbone access generates incentives for the IBP to reduce the quality of the backbone access sold to the local incumbent's subsidiary ISP.[24)]

Consumer demand for connection to ISP A falls if the quality of its backbone access has been reduced. Thereby the IBP will capture a larger share of consumer demand, and this is more profitable the larger the profit margin the IBP has in the end-user market. However, with a smaller consumer demand ISP A will also have a smaller demand for connection to the backbone. This harms the IBP, and this negative effect is more pronounced the higher the profit margin the IBP has on its sale to ISP A. Due to these conflicting effects, it is not obvious if and when the IBP finds it profitable to practise quality degradation.

In the following it will be convenient to denote by $\overline{w}_g$ the regulated price on backbone access, and by w^*_g the price that the IBP would have chosen in the absence of regulation. To avoid unnecessary notation we will assume that the IBP's own long-run marginal costs are equal to zero (this has no qualitative effects on the present discussion).

No regulation of the domestic market

Though the purpose of this section is to discuss the effect of two-sided regulation, it may nonetheless be instructive to first assume that only the IBP is price regulated. Suppose that $\overline{w}_g$ is set equal to zero. In that case the IBP does not make any profit from connecting ISP A to the backbone, and will therefore obviously have incentives to foreclose ISP A. By sufficiently degrading the quality of backbone connection for ISP A, the IBP thus becomes a monopolist in the end-user market. Obviously, it makes sense for the IBP to degrade the quality also if $\overline{w}_g$ is somewhat above zero, but as the price cap increases it becomes increasingly expensive for the IBP to forego the profits that it could have earned by selling connection to ISP A.[25)] This effect is particularly strong if the Internet Service Providers offer differentiated goods, and there is a large consumer segment that prefers the services offered by ISP A. In any case, the IBP will not have any incentives to degrade the quality if $\overline{w}_g$ is equal to w^*_g: it will then be better off by extracting profits by serving ISP A with backbone access (even though that implies competition in the end-user market). Actually, we may expect this effect to dominate also when $\overline{w}_g$ is somewhat below w^*_g, in which case the price cap will affect domestic welfare positively. But how will welfare be affected if $\overline{w}_g$ is set so low that quality degradation is profitable?

The first thing to note is that the integrated LAP need not be harmed by the quality degradation. This may seem a bit surprising, but the important point here is that the LAP controls the local bottleneck. Thus, if the IBP sets high prices in

---

[24)] *See Section 2 for a further discussion of the potential for quality degradation and why it may seem more relevant to discuss this point for IBPs than LAPs.*

[25)] *This also demonstrates an important point, namely that degradation is not a goal* per se*, but only a means to transfer market power from the regulated global bottleneck to the retail market.*

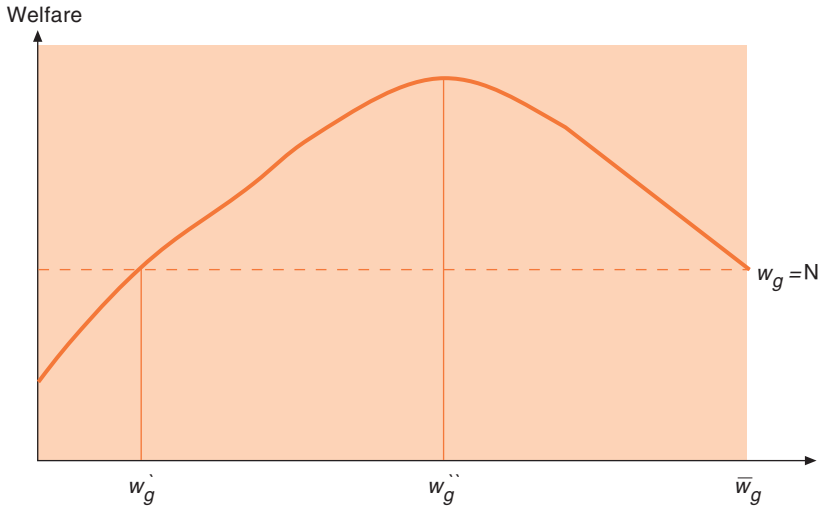


Figure 8 Welfare effects of a price cap on backbone access

the end-user market because it has become a monopolist – or at least the overwhelmingly dominating firm – the LAP can respond by setting high prices for local access. Thereby the LAP may be able to capture the larger share of the "monopoly" profit.[26] In other words, it may be more profitable to be a "bit-stream" wholesaler than to compete for consumers in the retail market. However, the higher consumer price implies that the deadweight loss increases.

Price Regulation of Both the LAP and the IBP

The optimal domestic regulation policy becomes quite complex if there is a price cap on backbone access and the IBP has the ability to practise quality degradation (or any other kind of non-price discrimination). However, it is immediately clear that the regulator will be unable to prevent quality degradation if $\bar{w}_g$ is close to zero, because the profit margin of selling backbone access to the integrated LAP is then very small. In that case the regulator should simply provide the integrated IBP with cheap access to the local loop, ie. set a low value of w_l, in order to reduce consumer prices. Likewise, it is clear that the regulator should set a relatively low value of w_l if $\bar{w}_g$ is close to w_g^*: quality degradation is then not profitable in any case, and the regulator should stimulate competition.

For some intermediate values of $\bar{w}_g$, which still are so low that there would be quality degradation if the LAP is unregulated, we may foresee a more surprising regulation policy. Suppose, namely, that the regulator sets a higher value of w_l than the one the LAP would choose. Then the profit margin $(p - w_l)$ for the IBP of selling to the end-user market will be relatively low, and it may therefore be better off by increasing the sales of backbone access to ISP A at the profit margin $\bar{w}_g$. We may thus end up with a regulated price that is higher than the one preferred by the LAP.[27]

Finally, Figure 8 shows how we may imagine that a price cap on backbone access affects domestic welfare. The broken line shows the welfare level when the IBP maximizes profit and charges $w_g = w_g^*$ (no price cap on backbone access), and the solid line shows the welfare level when the access price to the backbone is regulated.[28] The latter curve has an inverted U-shape, and lies partly below the broken line. The intuition for this is most easily understood if we consider the extreme values of w_g. At the far right we have $w_g = w_g^*$, and thus a non-binding regulation. Then welfare must necessarily be the same as in an unregulated economy. At the far left we have $\bar{w}_g = 0$, and the IBP unambiguously will foreclose ISP A completely. Thereby the IBP will use monopoly prices in the end-user market; consumer prices will be high, and the welfare level at a minimum (in particular, it will be lower than when $w_g = w_g^*$). Due to the income generating effects of selling backbone access to ISP A, it is less beneficial for the IBP to degrade the access quality the higher the price cap $\bar{w}_g$. Simultaneously, the *direct* effect of a binding price cap is to reduce consumer prices and access prices for ISP A (that is the reason for controlling the access price in the first place). At some point $w_g = w_g^{`}$ we should thus expect that welfare is higher than when w_g is not regulated. For sufficiently high values of w_g – at $w_g^{``}$ in the figure – it will no longer be profitable for the IBP to degrade the quality at all. This is the value of $\bar{w}_g$ that maximizes welfare; the only effects of increasing $\bar{w}_g$ beyond this value are to increase domestic prices and the profit level of the IBP.

---

[26] *Note that we have a clear symmetry between this result and the result that the IBP profits from a one-sided regulation of the local access price. In both cases the key point is that the unregulated firm controls an essential input that is needed by the other firm.*

[27] *Recall that the LAP may benefit from foreclosure due to the high consumer prices.*

[28] *The figure is drawn under the assumption that connection to ISP A and ISP B is imperfect substitutes for the consumers.*

The exact shape of the curve in Figure 8 should, of course, not be interpreted literally. The important point is that a price cap on backbone access should not be so low as to provoke socially wasteful foreclosure practises. That may turn out to be detrimental to welfare.

4 Conclusion

In this paper we have argued that a cost oriented regulation of the local access input may be inferior to the domestic government in an open economy. A one-sided price regulation that sets price equal to long-run marginal costs for local access may lead to increased profit shifting to foreign countries. Moreover, any international regulation policy should take into account the fact that dominating firms in the communication industry may impose socially wasteful non-price discrimination if their prices are too strictly regulated.

We have abstracted from network externalities in our analysis, even though these clearly are present in the Internet. Some readers may thus question the robustness of our results, since the network externalities possibly imply that firms will charge low prices even in an unregulated economy. Network externalities do not imply, however, that one should offer services for free. It may rather imply that the optimal price is lower than it would otherwise have been, and this only has quantitative effects. In this sense the above discussion does not depend on whether we consider network externalities or not. The fact that the IBPs have begun to charge ISPs for backbone connection may further indicate that the larger share of the network externalities in this industry already has materialized.

Dominant firms often argue that *ex ante* regulation will have detrimental effects, but in our view there may be strong arguments in favour of *ex ante* regulation of local access. The access price should not, however, be so low as to generate excessive profit shifting to foreign countries. The optimal policy may be that the regulator commits to set an access price above costs, possibly the same price as in an unregulated market economy.

References

Bailey, J, McKnight, L (eds.). 1997. *Internet Economics*. MIT Press.

Clark D. 1999a. Implications of Local Loop Technology for Future Industry Structure. In: Gillett, S E, Vogelsang, I (eds.). *Competition, Regulation and Convergence*. LEA, London.

Clark, D. 1999b. High-Speed Data Races Home. *Scientific American*, 1999. Available at http://www.sciam.com

Cremer, J, Rey, P, Tirole, J. 1999. *Connectivity in the Commercial Internet*. Toulouse, IDEI. (Technical Report.)

Economides, N. 1998a. The incentive for non-price discrimination by an input monopolist. *International Journal of Industrial Organization*, 16 (3), 271–284.

Economides, N. 1998b. *Raising Rivals' Costs in Complementary Goods Markets : LECs Entering into Long Distance and Microsoft Bundling Internet Explorer*. Stern School of Business, N.Y.U. (Discussion Paper EC 98-03.) (Available at www.raven.stern.nyu.edu/networks/)

Economides, N. 1998c. Competition and Vertical Integration in the Computing Industry. In: Eisenach, J A, Lenard, T M (eds.). *Competition, Innovation and the Role of Antitrust in the Digital Marketplace*. Kluwer Academic Publishers.

Economides, N, Salop, S C. 1992. Competition and Integration among Complements and Network Market Structure. *The Journal of Industrial Economics,* 60 (1), 105–123.

Foros, Ø, Kind, H J, Sørgard, L. 2000. *Access Pricing, Quality Degradation and Foreclosure in the Internet*. Bergen, Stiftelsen for samfunns- og næringslivsforskning. (SNF WP 12/2000.) (Available from www.snf.no)

Foros, Ø, Kind, H. 2000. National and Global Regulation of the Market for Internet Connectivity. Presented at the *3rd. Internet Economics Workshop,* Berlin, May 2000. (Available from www.berlecon.de/iew3)

Foros, Ø, Hansen, B. 1999a. *Competition and Compatibility among Internet Service Providers*. Presented at The Earie 26th Conference, Turin, 5–7 September 1999.

Foros, Ø, Hansen, B. 1999b. *Competing ISPs' incentives to increase interconnection quality*. Kjeller, Telenor R&D. (R&D Report R 27/99.)

Frieden, R. 1999. Last days of the free ride? The Consequences of settlement-based interconnection for the Internet. *The journal of policy, regulation and strategy for telecommunication, information and media,* 1 (3), 225–238.

Laffont, J-J, Tirole, J. 2000. *Competition in Telecommunication*. Cambridge, Massachusetts, The MIT Press.

Mackie-Mason, J, Varian, H. 1997. Economic FAQs About the Internet. In: Bailey, J, McKnight, L (eds.). *Internet Economics*. Massachusetts, MIT Press.

Milgrom, P, Mitchell, B, Srinagesh, P. 1999. Competitive Effects of Internet Peering Politics. Presented at the *27th Annual Telecommunications Policy Research Conference,* Washington DC, September 1999.

Mognes, P, Nord, T. 1999. Regulation of broadband access networks. *Telektronikk*, 95 (2/3), 50–59.

Shapiro, C, Varian, H. 1998. *Information Rules : A Strategic Guide to the Network Economy*. Boston, Massachusetts, Harvard Business School Press.

Spengler, J. 1950. Vertical Integration and Anti-Trust Policy. *Jounal of Political Economy,* 58, 347–352.

Speta, J B. 2000. Handicapping the Race for the Last Mile? A Critique of the Open Access Rules for Broadband Platforms. *Yale Journals on Regulation,* 17, 40–91.

Srinagesh, P. 1997. Internet Cost Structures and Interconnection Agreements. In: Bailey, J, McKnight, L (eds.). *Internet Economics*. Massachusetts, MIT Press.

Varian, H. 1998. How to Strengthen the Internet's Backbone. *Wall Street Journal*, 8 June.

Varian, H. 1999. *Market Structure in the Network Age*. Available from www.sims.berkeley.edu/\~hal/

Werbach, K. 1997. *Digital Tornado : the Internet and Telecommunications Policy*. Federal Communications Commission. (OPP WP 29/99.)

Pricing and Admissions Policies for IP Networks

JUDITH A. MOLKA-DANIELSEN AND KETIL DANIELSEN

Resource reservation systems are becoming an important tool as network providers are beginning to offer multimedia support, virtual leased lines and private networks over packet-switched internetworks. *Policy-controlled access* to the reservation system is necessary, but may fail to capture the immediate needs of the user. *Market-based access* control offers an alternative to policy-based control, where requests are admitted based on the current willingness to pay for reservation service. It is important to discover the level of value these market-based access controls offer to customer. Do these market-based mechanisms improve the customer perceived *quality of service (QoS)* over simple policy-based access such as first-come first-serve queuing? In this study, we examine several admission policies using market-based access controls. Our study seeks to attain whether value improvements over policy-based admission mechanisms could be obtained using market-based access control mechanisms such as auctions and reservation options. We have implemented a market-based admissions system that use bids and an auction-based admission rule to determine which customers gain access to network resources. We also have studied the use of a rule that prevents the preemption of network resources once they are gained.

Judith A. Molka-Danielsen (37) is Associate Professor with the Department of Informatics at Molde University College. She holds a Ph.D. (1998) in Telecommunications from the University of Pittsburgh, USA. She is teaching in the area of document and information management in enterprises, and the strategic use of information technology. Her research is in access and interactions policy issues in telecommunication markets.

j.molka-danielsen@himolde.no

Ketil Danielsen (34) is Associate Professor with the Department of Informatics at Molde University College. He holds a Ph.D. (1999) in Telecommunications from the University of Pittsburgh, USA. His teaching is in computer systems and network management. His research focuses on pricing and admission policy in telecommunication networks.

ketil.danielsen@himolde.no

Introduction

Internet protocols can be made to support usage pricing and admission control policies, and new policies on Internet Protocol (IP) networks are in early stages of testing and use. It is expected that Internet Services such as access, email, and Web hosting services are likely to be delivered over infrastructures which support other types of services such as telephone calls and video conferences. These services will have different network resource requirements and customers will have different expectations for quality of service (QoS) guarantees.

The default implementation of the ReSerVation setup Protocol (RSVP) in Internet routing technology begins to support different resource needs by enabling applications to signal per-flow requirements for reservations to the network (Braden 1997). Although there are concerns about the scalability of RSVP signalling on large networks, its per-flow signalling capabilities indicate that it is likely to have a place in the access portion of the IP networks. Among RSVP's limitations, the current implementation of the admissions policy without additions is not able to account for user value of network resources. In particular, RSVP allows for immediate reservations of bandwidth on a first-come, first-serve (FCFS) basis. When insufficient resources remain, requests are denied.

In this study we propose and test several hybrid resource allocation schemes that use a combination of policy controlled access and market based access control. The market based access control uses a pricing mechanism which can be supported by RSVP over IP networks. The admissions policy has the potential to give greater value to customers. A bid price can be added to the RSVP request message and an auction performed to determine fair market price for the resources and to decide which customer requests win access. This approach impacts end-to-end QoS because customers that win reservations in one time period can lose reservations prematurely in following auctions. To insure against losing access an alternative to hold the reservation would exclude customer and reserved resources from future auctions.

In particular, we report the results of an IP network simulation where we examine three types of service admissions models: FCFS, auction-based with preemption, and auction-based with an alternative to hold the reservation. This study is novel because we simulate these admissions control policies in RSVP over an IP network. We inquire if value improvements can be gained for customers.

QoS and Network Performance

Resource reservation systems attain to be an important tool for network providers that offer multimedia support, virtual leased lines and private networks over packet-switched internetworks. For these systems a definition of quality of service is needed. Quality of service can be characterized by welfare measures and network performance measures. Policy-controlled access methods such as administrative flow control and admissions control that regulate access to network resources attempt to improve consumer welfare and network performance. However, these administrative allocation mechanisms alone have failed to maximize the factors of con-

sumer welfare and network performance (Tanenbaum 1996).

The general problem is that network performance in IP networks will decrease under conditions of high offered traffic and longer transmission paths with more network segments. This can result in longer packet or message latency and in more packet and message loss on datagram packet switching networks like the Internet (depending on the transport protocol used: TCP, UDP).

- *Packet latency* can be the sum of delays such as propagation time, transmit time, queueing delay and processing time.

- *Packet loss* can occur when packets are dropped because of no more room in a queue. A service like TCP has guaranteed packet delivery. Because TCP re-sends unacknowledged packets, dropped packets result in increased packet latency since packets must be re-sent.

Policy control mechanisms like *flow control* can be used with TCP. The producer will refuse to send more packets when there is a certain number *n* of unacknowledged packets. This is to stop receiver queues from overflowing and is between source and destination. These mechanisms are pre-programmed into software and are not controllable by the consumer. Alternatively, *congestion control* protocols are aimed at maximizing the traffic that can be carried between aggregate sources. These protocols must often be combined with policies such as mandatory metering to limit offered traffic. Pricing policies can also be used with congestion control by offering higher tariffs for access during more congested times.

In our study, we simulate an auction-based admissions system. We allowed preemption of accessed resources under one run and prevented it under a second run. In a summary of our findings, there is a measurable difference in QoS. In the first run that protects the reservation from being dropped, we found a deterioration of network performance that was worse than in the scenario allowing preemption of service. Network performance was measured using factors such as *setup time*, the time it takes for a user to receive an end-to-end reservation; *uptime*, the amount of time the user has end-to-end resources in a transaction; and *preemptions*, the number of interruptions in service. It was observed that network performance decreased most under conditions of high traffic demand and longer transmission paths with more network segments. This was reflected in longer waiting times to receive initial end-to-end reservations, and shorter overall transaction uptimes for those reservations. The network performance is represented in the accumulated economic value to users. That is, we have accounted for accumulation of user value only when the end-to-end reservations were in place. A more detailed description of the performance metrics follows in the results.

Reservation System

This section presents an overview of the reservation system, including the services, internal protocol and services required from underlying layers. The protocol is modeled as an extension to the Reservation Setup Protocol (RSVP) used for reservation signalling in the integrated services Internet (Braden 1997).

Service Overview

The service objective is to give the end user real-time network performance control, using a QoS control agent as that proposed in (Danielsen 1997). The end user on the client side of a network flow uses this agent to set his *end-to-end* bid, which is attached to his reservation request for that flow. This end-to-end bid becomes the ceiling on how much the network can charge per time unit for a complete end-to-end reservation. A completed end-to-end reservation is one where all nodes along the path have admitted the request. Reservable resources on a link are allocated using an auction that gives the resource to the highest bidders. The reservation service is preemptible in the sense that a bidder may lose his reservation in subsequent auctions.

Standard Reservation Protocol

An RSVP session is initiated in the network when senders begin transmitting Path messages on a unicast or multicast network-level session. A Path message defines the sender characteristics and follows the data stream downstream towards receivers. Each node along this data path installs sender state based on the Path message. Upon receipt of the Path message, a receiver *i* may choose to send a Resv message containing a suitable resource reservation request for the announced sender characteristic. The Resv message includes the amount to reserve, $q(i)$ and travels upstream hop-by-hop towards the sender. An intermediate node evaluates the request and *immediately* queries the local admission control module for the link. In the default first-come, first-serve model, admission is granted and resources reserved by the admission control module if there are sufficient reservable resources for the new request. RSVP installs local reservation state for this request and forwards the request upstream, *only* if the request was admitted locally. Otherwise, the node returns a ResvErr message indicating the failure reason.

Nodes in the network (senders, receivers and intermediate routers) maintain soft sender and reservation state. Sender (and reservation) state must be refreshed by the periodic transmission of Path messages from the sender (and Resv messages by the receivers). If sender (or reservation) state in a node changes, the node immediately sends a refresh message downstream (or upstream). Sender (or receiver) state is removed in a node if the sender (or receiver) requests an explicit PathTear (or ResvTear).

A reservation is in place as long as the receiver keeps refreshing it. It is removed when the receiver explicitly removes it, or if it times out. Thus, there is no time limit on how long the reservation can last (nor how much the reservation is for), and a Tragedy of the Commons outcome may follow (Hardin 1968). During periods of network congestion, an economic model to price reservable resources could be used instead of the current FCFS admissions policy. The next section defines extensions to the basic reservation system.

Bidding Extensions

In a bidding service, end user i uses a personal QoS agent to specify $u(i)$, the user's maximum willingness to pay for a reservation. This QoS agent instructs the local reservation service to attach this bid to the user's flows. Admission control in each node will use a local auction to determine which requests to admit. The end-to-end bid must be split upon each of these auctions, so the receiver-side reservation service computes a *per-hop* bid $u_k(i)$, which is the ceiling on what node $k \in N_i$ (where N_i is the set of nodes that carry i's data) can charge for a link reservation in order to avoid end-to-end budget overrun, i.e. that

$$\sum_{k \in N_i} u_k(i) \leq u(i).$$

The per-hop bid is computed by the receiver, as the receiver's end-to-end bid divided by the number of hops between the sender and the receiver, i.e. $u_k(i) = u(i) / \# N_i, \forall k \in N_i$. The number of hops, $\# N_i$, is known by the receiver from the most recent Path message. The per-hop bid is included as a new object in the Resv message along with the standard reservation request. Since this design assumes the equal-bid bidding strategy, the per-hop bid is identical for each hop, and each node simply forwards the reservation request and bid unmodified. Bids in Resv messages are processed by the upstream nodes and stored as part of the link reservation state.

These simple unweighted per-hop bids are used because the reservation requests may be forwarded upstream before prices for a particular hop are computed. Alternative bidding strategies could be, for example, descending bids where the greater part of a reservation budget is placed on auctions that are encountered for initial hops. If these link auctions are won, the per-hop price $p_k(i)$ would be deducted from the bid budget, and the remaining budget could be used in auctions upstream; i.e. $u_k(i) = u_{k-1}(i) - p_{k-1}(i)$, assuming that node k is one hop upstream to node k–1.

This is impractical to implement because the auction for the first hop must be performed before the remaining bid can be sent upstream. Thus, to be able to forward the reservation request, admission must be performed immediately, but this could pose performance problems to a highly loaded reservation system. Another alternative bidding strategy could use bids that were weighted for a given hop based on the historical prices for that hop. This would require the receiver's agent to be smarter by maintaining historical information on per-hop prices. The chosen method does not use price information as feedback to optimize the receiver's per-hop bids.

Admission Algorithm

The admissions control algorithm must work well with the characteristics of the IP traffic. In particular the demand for reservations can be bursty and fluctuate greatly over short periods of time. In our simulation we select to use link auctions at fixed intervals which are periodic and independent of request arrivals to determine which requests to admit. While immediate admission control could easily overload the reservation system, periodic auctions create a more predictable environment for the clients.

The admissions algorithm is as follows as depicted in Figure 1. A new reservation request (denoted NEW) arrives in a Resv message which contains the amount of resources requested, bid price, hold reservation flag (on or off), and *end time* for the requested reservation, starting immediately. The node is currently serving a set of reservation requests A. If the new request can not be fitted it is included in A'. An auction is

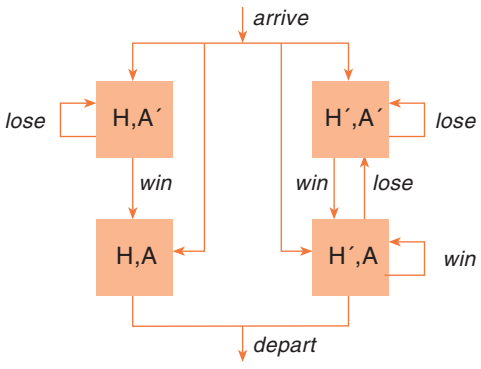


Figure 1 Request state diagram. State denotes whether the request is held or not, and if it is admitted or not. The arrows represent a request's arrival, departure, or win/lose events from being part of auctions

held periodically between requests in *A'* and in *H'* (those that are admitted and in *H*, keep their reservations). Winners are moved to *H* or *H'*, depending on whether they wanted hold option or not, while losers move to, or stay in *A'*.

The auction also determines a new spot market price that is to be paid by all admitted requests. This price is the second-best price. The price will be different for each held auction, since the marginally excluded requests are different for each winning request in a given auction. Since requests in *H* are guaranteed a ceiling in their bandwidth price, an insured request pays the minimum of the new spot market price and its original bid price which was given at the time when it was first admitted with insurance. In effect, if NEW wants to pay insurance, it pays its bandwidth bid price plus the current hold option price. Requests in *H'* will be paying the second-best price from the auction. The receiver decides, when placing the bid, to take the hold option or not. The holding price is computed at the node and this information is used later in the billing process. Notice that NEW may lose in the auction, and not be admitted, even though it is willing to pay hold option. It may of course resubmit its request later. Only after admission will NEW be guaranteed against preemption.

Auction Algorithm

The user bid is used by nodes in link auctions, if remaining network resources are insufficient to meet all current requests. This auction is modeled after the Generalized Vickrey auction, where bids are sealed and winners pay a second-best price (MacKie-Mason 1994). Users will theoretically then bid their true value. The second-best price is the bid price of the highest losing bids. The advantages are that signalling overheads are minimal, compared to the more commonly used English auction where the bidders use an interactive and signalling-intensive process to determine winners (Riley 1981).

A unit bid is computed for each request as $b_k(i) = u_k(i) / q(i)$, and used as a normalizing measure in ordering the requests before the auction, where the highest unit bids among the competing requests are chosen as long as there are resources left. Request *i* is to pay a unit price equal to the sum of unit values that all other requests would have gained *without i* present, minus the value they gain *with i* present. Let *i* be one of the admitted requests, and assume that the end-to-end personal valuation for the requesting user is $U(i)$. The theory for Generalized Vickrey Auction then says that the user bids an amount equal to his value, i.e. that $u(i) = U(i)$. In the adopted equal-bid strategy $u_k(i)$ is taken to be the user's per-hop valuation.

For the local auction, let $b_k(i,a)$ be *i*'s per-unit gain from an admitted set *a* in node *k*, x^* be the set of admitted requests after the actual auction where *i* was a winner, and $\hat{x}_{-1}$ be the set of admitted requests if *i* had *not* been part of the competing requests. The per-unit spot price $p_k(i)$ for an admitted request *i* is computed as for the Generalized Vickrey Auction:

$$p_k(i) = \sum_{j \neq i} b_k\left(j, \hat{x}_{-1}\right) - \sum_{j \neq i} b_k\left(j, x^*\right)$$

Note that $b_k(j,x)$ equals *j*'s unit bid if *j* is one of the admitted requests in *x*. The spot market price, $p_k(i)$, is used to compute the request's usage payment for the current period, which becomes $tp_k(i)q(i)$, where *t* is the accounting period length and $q(i)$ the quantity reserved by *i*.

Related Work

Users of the basic reservation service are likely to see service oscillations during high demand for reservable resources, since an established end-to-end reservation may go on and off as nodes asynchronously preempt and admit. A reservation holding service option has been proposed in which the end user can choose to purchase a hold contract and be guaranteed against service interruption. Our study was inspired by another study conducted by Lazar and Semret using a hold option for access to indivisible modem lines in a modem service (Lazar 1999, 2000). Their study differs from ours, in that bandwidth is a divisible service in the dimensions of quantity and duration of usage. These authors also develop the concept of hold options as a market mechanism. In particular, their patent pending model evaluated a modem service with a limited number of lines that allows users to bid per time unit (minute) for access to a modem line for a given duration. Price for receiving service is a per-time unit spot price and a hold option price that insures availability for the entire time interval. Using real traffic traces they obtained significant value improvements from the modem server.

Reservation System Model

A simulation model of a market-based reservation system has been built using the *network simulator* environment (Fall 1998) with modified RSVP extensions (Greis 1998). The network topology studied is a model of a symmetric 80-node network with 5 backbone nodes, 15 access nodes and 60 end nodes, as illustrated in Figure 2. All links are full duplex and point-to-point.

Minimum and maximum number of hops that a request traverses are two (within access area)

and six (worst-case has two backbone hops). For a given node there is a fixed number N of sending users that independently initiate 64 kb/s unidirectional flows towards randomly selected end nodes (uniform between the 59 other end nodes). At most, $60N$ flows may be active within the system. The probability that a new flow (towards a randomly selected receiver) traversing 2, 4, 5 or 6 hops are 3/59, 8/59, 24/59 and 24/59.

Each session has a random duration (exponential with mean 60 seconds), after which the sender waits a random idle time (exponential with mean 30 seconds) before initiating the next session as described. The receiving user replies to the first Path message with a Resv message for the whole duration, specifying the amount (64 kb/s) and a random per-minute bid (uniform between 5 and 10) that remains unchanged for the session. The user bid is divided evenly on each link along the path. This means that a user bidding 8 for a 4-hop path places a bid of 2 for each link. Auctions are performed independently on each link, and if batch admission is used, on average every $T = 5$ seconds.

Performance Metrics

Both economic and engineering performance is of interest in the study and is described as follows. First, however, note that a flow's reservation is said to be *complete* if all reservation requests on a path from the sender to the receiver have been admitted. Economic performance metrics are

- **Consumer value.** This is the receiver's average value (assumed to be equal to his bid) accumulated only while the reservation is *complete*. No value is generated while one or more of the nodes fail to admit the reservation request.

- **Usage payment.** This is the average receiver's payment for reservation service, and is (as for value) only accumulated while the reservation is complete. Payment is the sum of spot prices resulting from local auctions in each node. Payments are not accumulating while one or more nodes fail to admit the reservation request.

Consumer surplus is the difference between consumer value and usage payment, and is also reported on. Engineering performance metrics are:

- **Average setup time.** This is the average time it takes to reach the *first* completed reservation for a flow, where all reservation requests for the flow have been admitted.

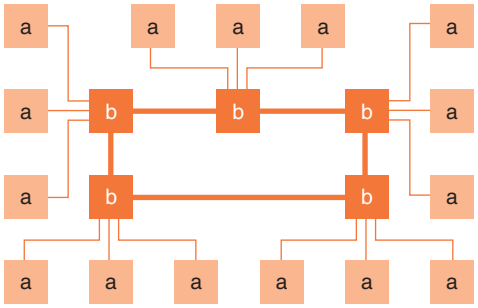


Figure 2 Simulation topology. Backbone nodes are marked 'a' and interconnected as a ring with 1 Mb/s links. Access nodes are marked 'a' and connects to one backbone via a 0.5 Mb/s link. End nodes are not shown but there are four end nodes, each attaching to its access node via a 0.3 Mb/s link. Each end node has an adjustable number N of senders generating flow workload

- **Completion ratio.** This is the average share of initiated flows that gained one or more complete reservations.

- **Uptime.** This is the average share of a flow's duration, in which the reservation is completed, computed as the sum of individual completion durations divided by the flow duration.

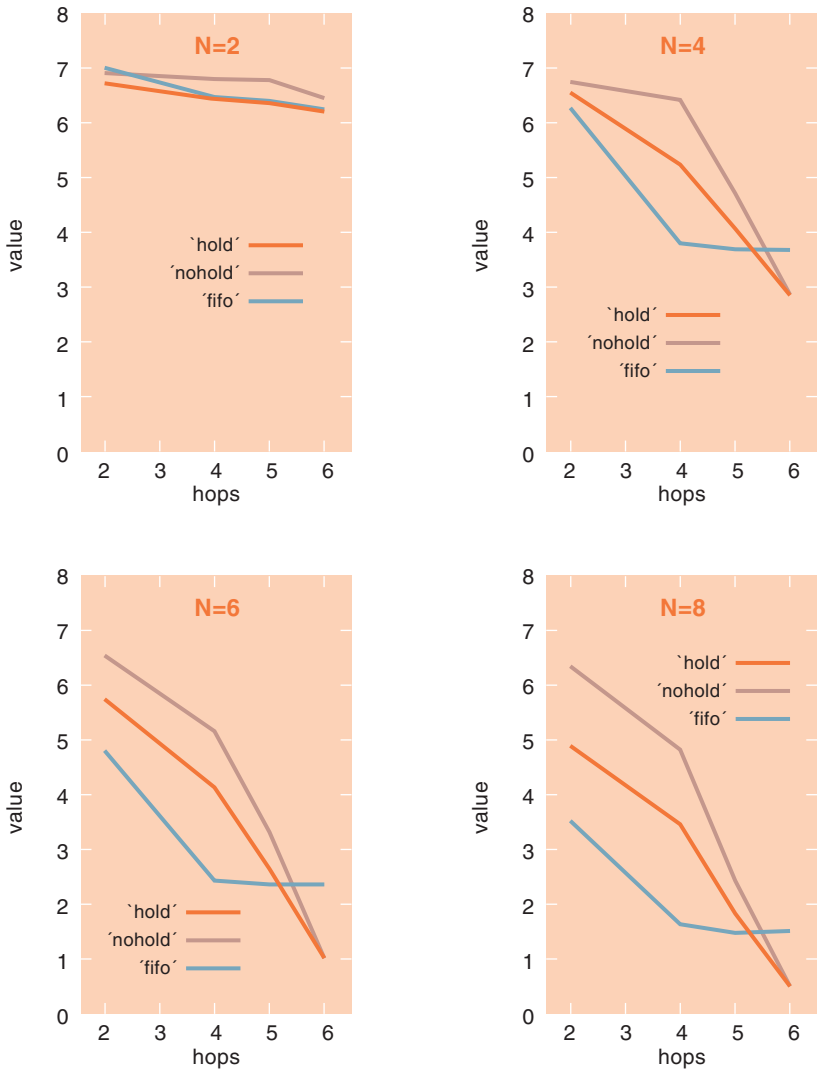


Figure 3 Average consumer value generated per flow, for loads N = 2, 4, 6 and 8, respectively

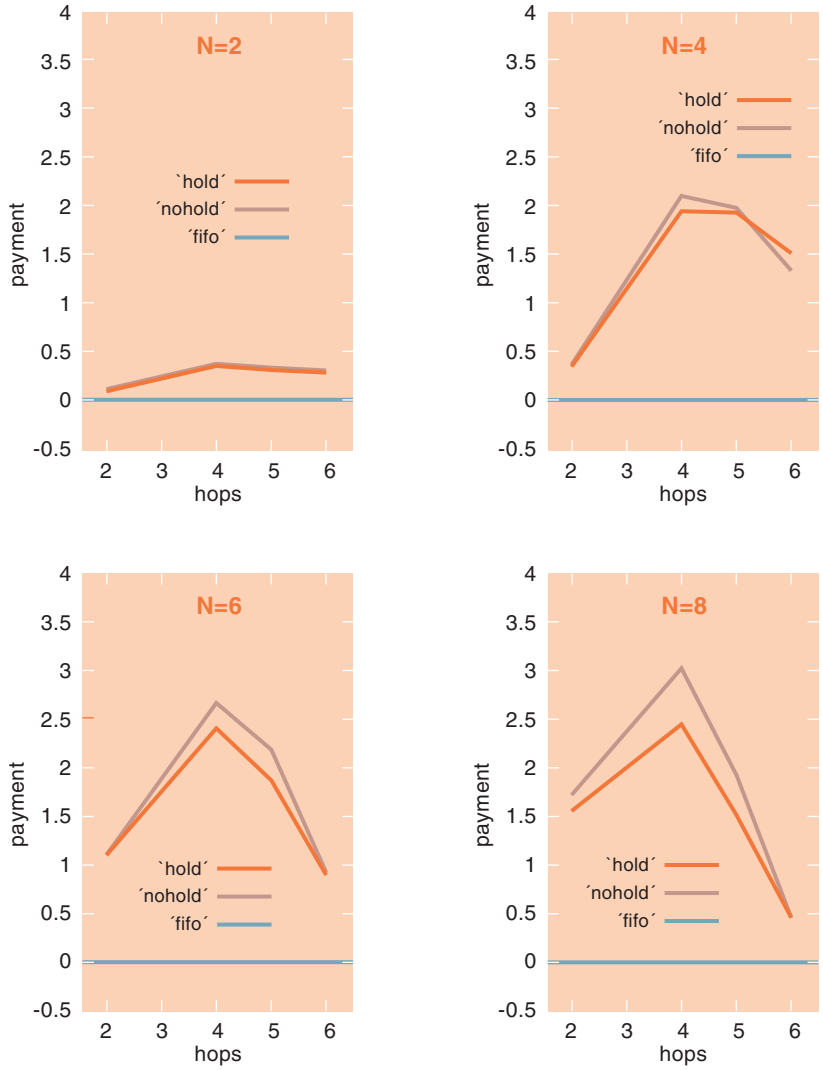


Figure 4 Average consumer payment generated per flow, for loads N = 2, 4, 6 and 8, respectively

- **Preemption ratio.** This is the average preemption count, divided by the total number of flows. It is a measure of how often completed flows go down and not a count of how many flows were preempted.

All means are the sum of all samples divided by the number of observations. Note that all averages are per-flow, and that variance probably can be attributed to the large difference in flow durations (which were exponentially distributed with a mean of 60 seconds).

Simulation Results

In this section we summarize the results from the simulation runs (Molka-Danielsen 1999). We refer to three types of runs. Runs using auctions with a hold on reservations are refered to as Type I. Runs using auctions without hold on reservations are refered to as Type II. Runs using FIFO queuing are referred to as Type III. System load, measured in the number of senders per node were $N = 2$, 4, 6, and 8, which gives a maximum number of simultaneously active flows of 120, 240, 360 and 480, respectively. Results and trends for these runs are presented in the order that they were described in the previous section. These are: consumer value, usage payment, consumer surplus, average setup time, completion ratio, uptime and preemption ratio. The economic performance metrics are summarize in Figures 3, 4 and 5. The engineering performance metrics are summarized in Tables 1 and 2. All reported averages are the arithmetic average of samples from six independent runs. Each run generated a sample based on the arithmetic average of approximately 3100 flows.

Observations

Our results reflect the delays found in networks with similar processing conditions (Stallings 1997).

- For *circuit switching networks* there is call setup delay. A call request must be sent through the network to see if the destination station is not busy, in which case a call-accepted signal is returned. A processing delay exists at each switching node when the call is setup, but not during the active call transmission. This is comparable to our study where the market-based mechanism is used to gain reservation access and the reservations are held. Held reservations reduce the availability of remaining network resources.

- In a *virtual-circuit packet switching network*, a virtual circuit is requested using a call-request packet at each switching node. But there are also delays in the transmission because the packets are queued at switching nodes to await transmissions. Delay is variable and increases with network load. This is comparable to our study where the market-based mechanism is used to gain reservations but the reservations are not held and end-to-end access must be re-established.

- In a *datagram packet switching network* there is no call-setup. For short messages or transactions, this kind of network can perform better than circuit switching networks. FCFS policies are used to gain access by default. Alternatively other pricing based access policies can be used. In FCFS access, when loads are low and messages are short, there is little delay in gaining access and low probability of interruption of service. This is comparable in our study to the policy controlled access using an FCFS rule to gain reservation access.

Economic Performance

Consumer value in all Types of runs decreases as the number of hops or length of a connection increases. With Type III (FIFO), values decrease between 2 and 4 hops but remain at nearly the same level for 5 and 6 hop connections. The consumer value also decreases for all types of

runs with increasing loads. Comparing between the types of runs, Type II (no-hold) had consistently higher consumer value than Type I (hold) as measured across all levels of load and number of hops. Type III (FIFO) had lower consumer value than Type I and Type II when the path length was 2, 4 and 5 hops. But Type III consumer value was greater than Type I and II for 6 hop connections.

Usage payments are related to the uptime per connection and the spot or strike price. There are no payments for Type III runs. For Type I and Type II runs the payments follow the same trends. Payments at first increase as the load increases for both 2 and 4 hop connections. But where the number of hops is 5 and 6, the payments begin to decrease again under high load conditions where there are 6 and 8 flows per node. This is because payments are only activated for completed reservations where all nodes have admitted the reservation request, and it is more difficult to obtain completion under higher load because there are less available resources. It was also observed that there were small differences in payments where Type II (no-hold) payments were only slightly higher than Type I (hold) payments. We see this outcome as reasonable. In the Type II (no-hold) scenario the consumer pays the spot price which is always lower than their bid. If the spot price is above their bid, they lose the reservation and they do not pay. Payment accounting is then stopped. In comparison, in Type I (hold) runs the consumer pays either the spot price in periods where spot price is below their bid or they pay the strike price (equalling their bid price) in periods where spot price is above their bid. They are paying for additional periods, but it also takes longer to set up the connection, as is shown in the setup times. So, accounting begins later.

Based on the trends for consumer value and usage payments described above the trends for consumer surplus followed those values. In summary, consumer surplus for Type I (hold) runs decreases with an increasing number of hops in the connection. It also decreases with increasing load. Type II (no-hold) runs followed the same trends as Type I runs. Comparing Type I to Type II runs, across variables of hop length and load, the Type II (no-hold) runs in all comparisons had a consumer surplus that was greater than in Type I. Comparing Type II (no-hold) to Type III (FIFO), Type II had a greater consumer surplus for shorter paths (2 and 4 hops) while the FIFO performance flattened out and was better under longer length connections (5 and 6 hops). Note again there are no usage payments in FIFO, so it was not surprising that consumer surplus for Type I and II runs drops below FIFO under conditions where payments become large.

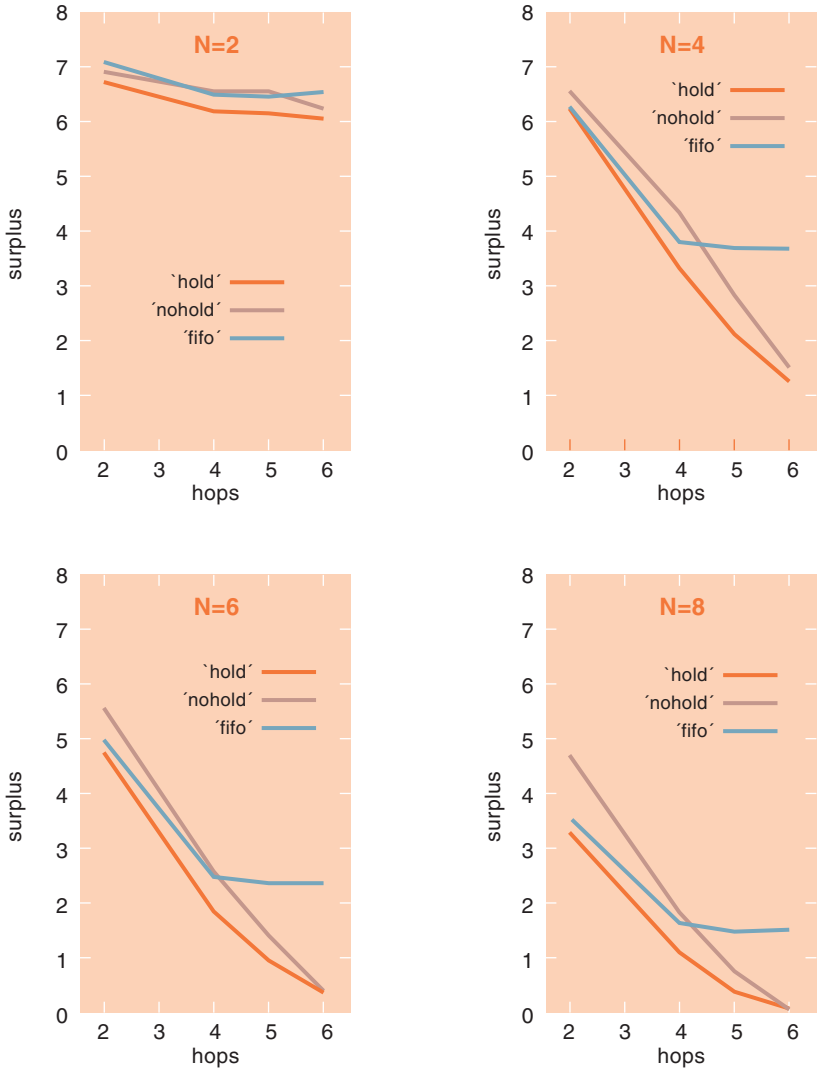


Figure 5 Average consumer surplus generated per flow, for loads N = 2, 4, 6 and 8, respectively

Engineering Performance

The engineering performance metrics can help to further explain some of the results in the economic performance metrics. In each comparison, the average setup times were longer for Type I (hold) runs than for Type II. The setup times increased with longer hop reservations and increased with higher loaded networks. Type III runs had slightly longer setup times over Type I and Type II runs for 2 hop reservations. Type III runs had much longer setup times for connections of 4, 5 and 6 hops. They also had increasing setup times with increasing load.

The completion ratio for those flows that eventually achieved an end-to-end reservation was consistently higher for Type II over Type I runs. For loads of 2 flows per node the Type III completion ratio was not significantly different from Type I. For loads of 4 flows Type I completion ratio was higher than Type III. But for loads of 6 and 8 flows, Type III had a better completion ratio than Type I for longer connections of 5 or 6 hops.

Hop count	Run type	Completion ratio				Setup time				Uptime			
		2	4	6	8	2	4	6	8	2	4	6	8
2	I	.938	.902	.793	.735	3.7	5.5	10.9	15.5	.797	.750	.600	.505
	II	.947	.938	.953	.948	3.3	3.3	3.3	3.4	.818	.808	.817	.815
	III	.937	.870	.722	.527	3.8	6.7	6.9	7.0	.807	.718	.503	.300
4	I	.902	.708	.567	.473	5.7	17.1	22.7	27.2	.735	.483	.358	.278
	II	.935	.892	.782	.688	4.1	5.2	6.6	6.0	.785	.702	.588	.527
	III	.892	.523	.337	.257	6.7	32.8	57.7	75.1	.717	.287	.142	.088
5	I	.883	.537	.347	.238	6.3	24.6	30.0	35.1	.712	.337	.200	.130
	II	.922	.710	.517	.400	4.7	9.1	9.5	9.5	.755	.492	.358	.270
	III	.883	.520	.332	.235	6.9	34.9	59.7	61.9	.705	.273	.135	.078
6	I	.868	.348	.150	.070	7.2	37.5	46.1	52.7	.690	.193	.073	.033
	II	.900	.478	.223	.115	5.6	13.2	16.9	21.4	.697	.280	.117	.047
	III	.878	.508	.318	.220	7.0	36.7	61.9	80.5	.705	.258	.127	.073

Table 1 Engineering results for loads N = 2, 4, 6 and 8, and for hopcounts 2, 4, 5 and 6

The amount of uptime was highest for Type II (no-hold) connections. The reason the uptime for Type II is greater than that for Type I is that the setup time to first achieve the end-to-end connection was much longer for Type I than for Type II runs, even though the Type I reservations would be held. In an alternative design, if the average length of the duration of a reservation was to be made longer, then the impact of the length of the setup time might be reduced. The uptime for FIFO was less than it was for Type I and Type II. The uptime for FIFO was considerably less than the other types for connections of longer than 2 hops. Finally, the preemption ratio was measured and it only applies to Type II (no-hold) runs. The ratio became higher with longer hop reservations. The preemption ratio peaks at a load of 4 or 6. It is related to the completion ratio. If the number of completions goes down then the possibility for preemptions is also reduced.

Hop	2	4	6	8
2	0.000	.000	.000	.000
4	0.015	.167	.180	.140
5	0.093	.300	.213	.178
6	0.237	.352	.205	.128

Table 2 Preemption ratio for Type II (no-hold) admission, for loads N = 2, 4, 6 and 8, and for hopcounts 2, 4, 5 and 6

Conclusions and Summary

The results of this study are sensitive to several issues. First we have modified the implementation of RSVP to allow for the implementation of auctions and holds. As a result, the performance measurements in this study do not reflect the signalling performance of the current Internet proposed standard version of RSVP (Braden 1997). Second, we used traffic flow assumptions where there was an equal chance of a flow going to any receiver. Given the network depicted in Figure 2, this flow assumption has the impact of making the backbone resources busier than the end node resources. The traffic patterns therefore may not reflect real network traffic patterns. For example, the local links could carry more flows than the backbone links. Third, we assumed bids for resources were spread out evenly over all hops. This left fewer resources to win the busy backbone links.

Given stated qualifications, this study has some fascinating results for consumer value and quality of service. First the mechanism of the hold option has some negative effects when measuring the consumer value of network resources. That is, by allowing consumers to hold resources from future auctions, those resources are used at a lower bandwidth rate than could be obtainable in future auctions. As a result the average accumulated network value for the resource is lower when holds are used than when they are not used. In this study all or none of the users request the hold (Type I or Type II runs).

Second, in the runs where preemption is not allowed, it takes longer to establish a completed reservation as the number of hops in the connec-

tion increases and as the load on the network increases. Longer setup times are again the result of fewer resources being available to remaining players. The long setup times in the hold runs also have an impact on the overall uptime of the reservation where again the no-hold runs had longer uptimes than the hold runs. Available resources were also important for quickly regaining resources after a preemption.

In summary, the service alternative of a hold option should offer a quality improving alternative to customers. But, in our study the value gains to those guaranteed non-preemptible service were not realized. We think if a smaller percentage of network customers selected to use the hold the negative effects of tied-up resources could be greatly reduced, and there could be added consumer value to hold customers. Therefore, we suggest future studies should include a feedback mechanism based on volatility of price where the user includes the hold value with their bid and the network determines which bids are high enough to be given the hold option.

Acknowledgement

The authors wish to thank Marc Greis at the University of Bonn, for generously making his RSVP modifications and large-topology programs for the ns simulator available for this project (Greis 1998).

Bibliography

Braden, B et al. Resource ReSerVation Protocol (RSVP) – version 1 functional specification. In: *Internet RFC 2205*, September 1997.

Danielsen, K, Weiss, M. User control modes and allocation in IP. In: *Internet Economics.* Lee McKnight and Joseph Bailey (eds.). Cambridge MA, The MIT Press, 1997, 305–321.

Fall, K, Varadhan, K. *ns notes and documentation.* Technical report, Univ. of California at Berkeley, 1998. (2000, June 29) [online] –URL: http://www-mash.cs.berkeley.edu/ns.

Greis, M. RSVP/ns : An implementation of RSVP for the Network Simulator ns-2. In: *Technical report.* Computer Science Dept-IV, University of Bonn, 1998.

Hardin, G. The tragedy of the commons. *Science*, 162, 1243–1248, 1968.

Hull, J C. *Options, Futures and other Derivatives, third edition.* Englewoods Cliffs, NJ, Prentice-Hall, 1997.

Lazar, Aurel, Semret, Nemo. Design and analysis of the progressive second price auction of network bandwidth sharing. *Telecommunications Systems,* 2000. (To appear.)

Lazar, Aurel, Semret, Nemo. Spot and derivative markets for admission control and pricing in connection-oriented networks. In: *16th International Teletraffic Congress,* Edinburgh, UK, June 7–11, 1999.

MacKie-Mason, J K, Varian, H R. Generalized Vickrey Auctions. In: *Technical Report.* University of Michigan, July 1994. (2000, June 29) [online] – URL: http://www-personal.umich.edu/~jmm/research.html

Molka-Danielsen, J A, Danielsen, K. Pricing in admissions policy : Using hold options over ip networks. In: *Proceedings of the 1999 Telecommunications Policy Research Conference,* Washington DC, September 1999, Section 1, 1–10.

Riley, J G, Samuelsen, W F. Optimal auctions. *American Econimic Review,* 71, 381–392, 1981.

Stallings, Williams. *Data and Computer Communications, fifth edition.* Upper Saddle River, NJ, Prentice-Hall, 1997.

Tanenbaum, A S. *Computer Networks, third edition.* Upper Saddle River, NJ, Prentice-Hall, 1996.

Further reading

Related project information is available from http://www.himolde.no/~molka/ippricing.

Special

Managing QoS in Multi-Provider Environment – a Framework and Further Challenges[1)]

TERJE JENSEN, IRENA GRGIC, OLA ESPVIK AND METTE RØHNE

1 Introduction

Finding appropriate solutions for managing Quality of Service (QoS) becomes a growing challenge on the background of the steadily increasing number and types of actors (providers/users) and services offered by/to them. The challenge is further fuelled by the volatile commercial configurations.

In addition to those concerns, technical factors also contribute to the complexity in terms of more systems, protocols and mechanisms. Basically, a provider has to decide upon the service portfolio and the accompanying QoS and price levels. This implies that QoS belongs to the list of issues on the agenda for the business discussion. In fact, QoS-related activities may both become a main differentiation factor and be attached with fairly high cost.

These observations strengthen the need for achieving a more fundamental understanding of QoS-related issues. Such understanding can be built in a common generic QoS framework that would support different QoS arrangements in emerging complex situations mentioned above. A generic QoS-related framework has to be generally applicable for all types of users, providers, services, networks, etc. It would be rather cumbersome to relate oneself to various interpretations and principles for QoS depending on which services and systems are involved. Having a generic QoS framework alleviates the tasks of managing QoS in a multi-provider environment as it harmonises the understanding of QoS. Application of the framework would also enable more efficient agreeing on the QoS between actors involved in the provision of the service.

A generic QoS framework elaborated in the EURESCOM Project P806-GI (ref. [1], [2]) is described in Chapter 2, where its main concepts and terminology are introduced. In addition, a pivotal task of managing QoS in a multi-provider environment – agreeing upon QoS – is addressed by structuring input data and results. The concordance of the QoS framework with results from international bodies is addressed in Chapter 3, showing how those results are united in a common view. Selected open issues related to QoS in multi-service multi-provider environment are discussed in Chapter 4. Such discussions may motivate further studies undertaken by different groups and forums.

2 Agreeing QoS

2.1 Service Provision Challenge

Quality of Service (QoS) is a term that has been used for a long time and given various interpretations. Here it is defined as *the degree of conformance of the service delivered to a user by a provider, with an agreement between them* ([1]). The agreement is related to this service, and it represents a harmonised understanding between aforementioned entities, which is given as a set of statements. This means that QoS characterises the way the service is provided. It also means that the service characteristics (i.e. functions that are realised) are not considered as describing its QoS.

Ensuring a successful service delivery for requests originated in one provider's domain and involving other providers' domains, is a central issue. When a provider agrees some target QoS with a user, this (primary) provider may have to rely on service(s) provided by its sub-providers. Subsequently, the QoS delivered to the user may depend on the QoS delivered to the primary provider by the sub-providers. A provider has then to define and operate mechanisms for managing the QoS for its users and for other relevant providers. This involves both harmonised understanding of QoS terms and agreeing upon the QoS objectives, as well as monitoring/controlling whether these are obtained or not. In addition, as service degradations are likely to occur, adequate reaction procedures should also be considered and incorporated in the agreements.

Such service provision configurations may be depicted as depends-on graphs (see Figure 1), where a number of entities could contribute to a successful service provision to a user. Interfaces would be present between the entities. Services on various functional layers (transport services, directory service, management services, etc.) may be involved. Some of the entities involved could be composed of software only, implying that some of the interfaces may be of more logical nature.

Terje Jensen (38) is Research Manager at Telenor R&D, Kjeller, responsible for co-ordinating projects in the area of QoS and network design. He earned his PhD degree in 1995 from the Norwegian University of Science and Technology. Other activities include performance modelling and analysis, dimensioning and network evolution studies. He was Task Leader in EURESCOM P806-GI.

terje.jensen1@telenor.com

Irena Grgic (29) is Research Scientist at Telenor R&D, Kjeller. She is mainly involved in activities related to QoS and charging for different networks and systems, and studies related to network evolution, both in international and national projects. She holds an MSc in Electrical Engineering from the University of Zagreb in 1999. She was Task Leader of Task 5 in EURESCOM P806-GI.

irena.grgic@telenor.com

[1)] *This paper is based on a presentation made at Nordic Teletraffic Seminar no. 15.*

Ola Espvik (57) is Senior Research Scientist and Editor of Telektronikk. *He has been with Telenor R&D since 1970 doing research in traffic engineering, simulation, reliability and measurements. His present research focus is on Quality of Service and Quality Assurance mainly related to various EURESCOM projects. He holds an MSc in physics from the Norwegian University of Science and Technology 1968. He was Project Leader of EURESCOM 806-GI.*

ola.espvik@telenor.com

Mette Røhne (34) is Research Scientist at Telenor R&D, Kjeller. Her main activities include applied QoS, network design and techno-economic studies, performed both in international and national projects. She received her PhD degree in 1999 from the Norwegian University of Science and Technology.

mette.rohne@telenor.com

The fact that the set of business relationships might differ from the set of technical relationships adds complexity to the service provision. For instance, two actors may have a business relationship (e.g. involving payments), while the technical interfaces and the traffic flows pass through other actors. Therefore, both business and technical matters have to be kept in mind. Solving the above issues is simplified by applying an agreed QoS framework to the service.

2.2 A Procedure for Agreeing QoS

The procedure for agreeing QoS addresses how to arrive at a proper service provision configuration. However, this is not dealt with in every detail in this paper, as several business matters should be included. The process of considering QoS should be closely related to the business decisions made for arriving at the service delivery configuration. The inputs necessary to run the procedure (ref. Figure 2) can be categorised as follows:

- Description of the service provided by the primary provider to the user, e.g. in terms of user requirements;
- Types and value ranges of the QoS parameters as basis for negotiations between the user and the primary provider;
- The knowledge of potential business/technical configurations as seen from the primary provider enabling it to provide the service to the user, and
- Other relevant inputs, e.g. any regulatory concerns, financial issues, and so forth.

A service description commonly includes information on the entities involved. Information on the process of delivery of the service might be included. Also, all relevant interaction points should be identified, e.g. service access points, measurement points, observation points, (re)negotiation points, etc. Additional information (e.g. user specific requirements, usage conditions and constraints) might be attached.

Service provision configuration involves roles of business entities, as well as their relationships. A business entity (e.g. a company) is named an actor when it takes actions on the market. An actor could take on one or several roles. Besides, a number of actors could as well take on the same set of roles, e.g. in the case when competitors are present in the same market. Simply, there is not a single unique mapping between the roles and the possible business configurations/ scenarios. The relations between the roles may be both business and technical. Business relations comprise activities between different actors regarding e.g. legal and economic matters. On the other hand, technical interactions reflect activities between systems/networks regarding e.g. measurements, performance reporting, information transport, etc.

The objective of the procedure is to come up with the content of all QoS agreements and their relations relevant for the primary provider. The result of executing the procedure would then be:

- The description of the business configuration and the technical configuration as seen from the primary provider;
- The description of the agreement (business/ technical) between the user and the primary provider;
- The description of the set of agreements (business/technical) between the primary provider and any other provider involved that the primary provider has selected;
- Relationships between the above agreements, also implying that the (internal) mechanisms in the primary provider are settled.

When deciding which means to apply in order to reach consensus on the agreement, various tactical concerns are taken including both economic and technical issues. For example, some QoS aspects can be met by balancing internal means against the conditions seen in the agreements from sub-providers. The set of criteria to use

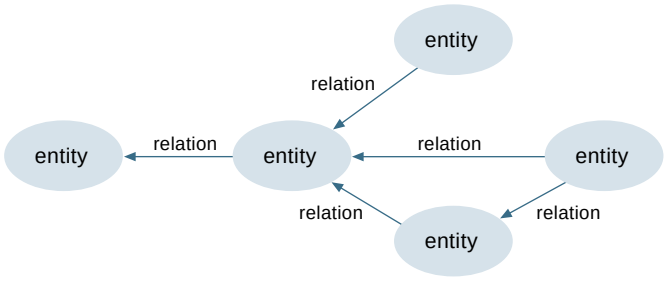


Figure 1 Multi-provider configuration may be similar to a depends-on graph

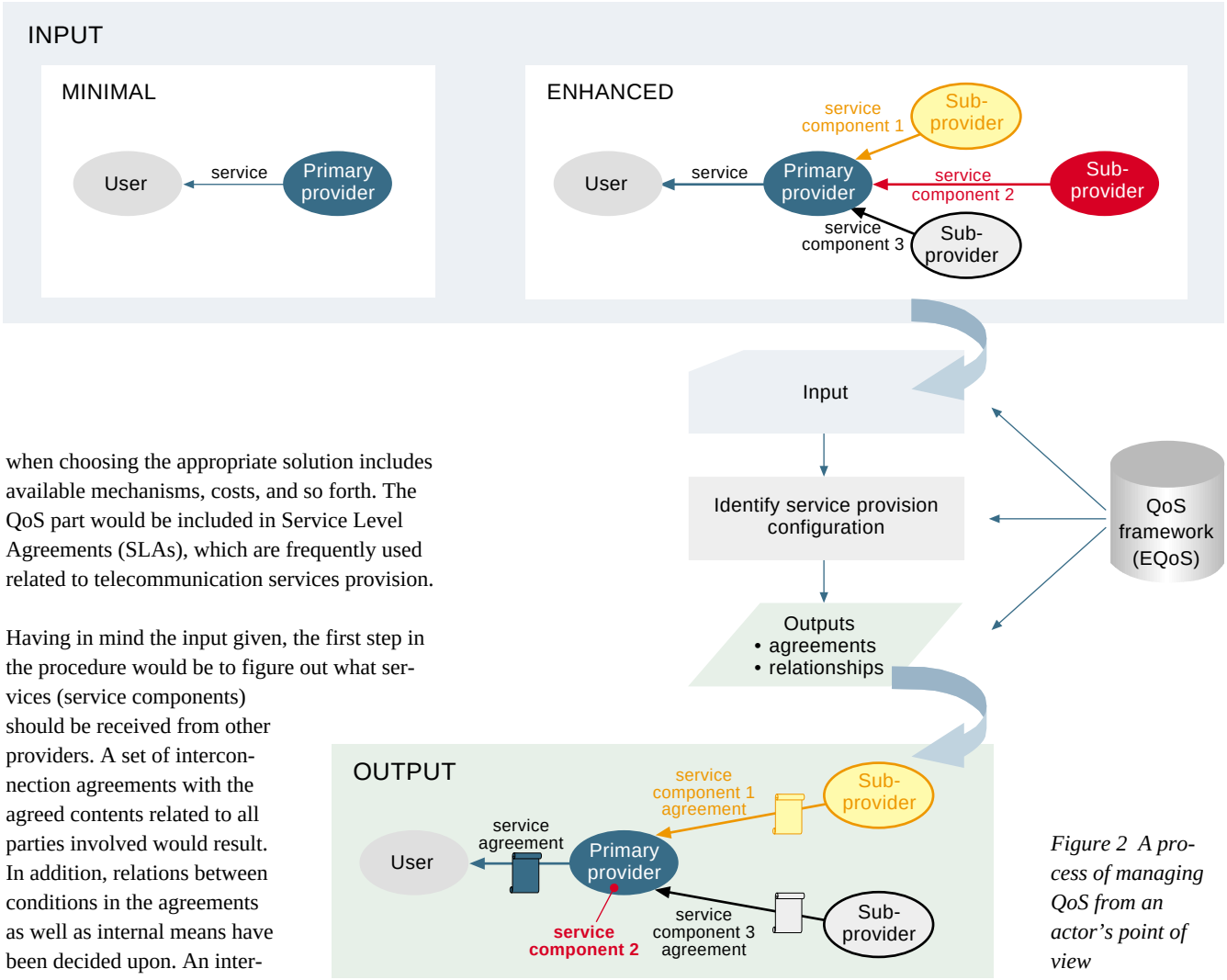


Figure 2 A process of managing QoS from an actor's point of view

when choosing the appropriate solution includes available mechanisms, costs, and so forth. The QoS part would be included in Service Level Agreements (SLAs), which are frequently used related to telecommunication services provision.

Having in mind the input given, the first step in the procedure would be to figure out what services (service components) should be received from other providers. A set of interconnection agreements with the agreed contents related to all parties involved would result. In addition, relations between conditions in the agreements as well as internal means have been decided upon. An interconnection agreement between a user and a provider should include the description of relevant both business and technical interaction points. Having identified the set of agreements implies that the service provision configuration has been agreed. Furthermore, the relationships between the agreements as well as internal mechanisms in the primary provider have been decided upon.

Generally, an interconnection agreement may include several issues, e.g. see [3]. A structure of the QoS-related parts of such an agreement is described in the subsequent section.

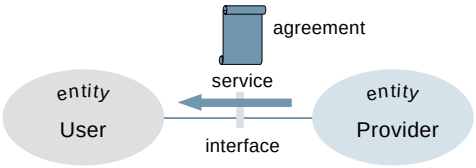
2.3 Generic QoS Framework

As mentioned above, deciding upon the QoS would be simplified by having established a generic QoS framework. In general, a QoS framework defines QoS-related terminology and concepts helping us to handle QoS. The EQoS framework brings a harmonised understanding of terms essential to managing QoS. It also introduces the notion of "one-stop responsibility" and elaborates the possibilities of applying it in a multi-provider environment.

The terminology achieved and incorporated into the EQoS framework would enable having unambiguous understanding of both service provision and QoS-related terms. The QoS term is defined in Section 2.1. Other terms related to service provision are explained in following and illustrated in Figure 3.

An *entity* is a generic unit characterised by its set of states and transitions. A number of entities can be composed into a new entity. Here, the behaviour of an entity is described as seen from an outside observer. It should be noted that the term entity relates to user or provider while the term actor refers to units related to service provision (e.g. organisation or person). An *interaction point* is a point where two entities exchange information. A set of interaction points constitutes a logical boundary between two entities i.e. an interface. An interaction point could potentially refer to a set of physical points (e.g. when referring to transfer delay between two physical points). A *service* is the result of executing a set of functions and is provided at the interface. It might be composed of multiple information exchanges between the entities. A *provider* is an entity that provides a service to another entity. A *user* is an entity that makes use of a service provided by another entity.

Figure 3 Agreement, encompassing descriptions of provider, user, service and interface

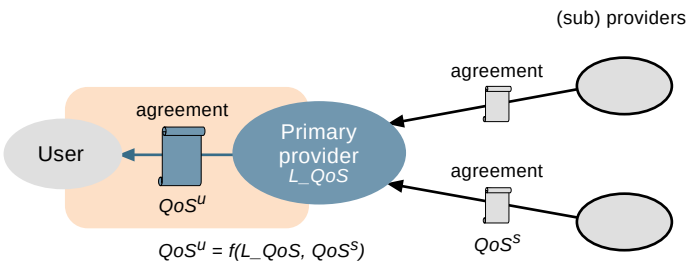


Such a fairly abstract level of description allows for capturing the provision of any telecommunication service. Furthermore, it is not restricted only to telecommunication services, but in a similar way it applies to provision of most services in other areas/industries. Thus, contributions from other areas can also be considered and appropriate results incorporated.

It is essential to observe that the quality of a service is, basically, different from the functions provided by the service. This means that QoS is not an attribute of the service itself, but it is rather related to how the service is provided. This recognises the choices facing a provider that typically include: (a) the service portfolio to provide, (b) the price of the services in the portfolio and (c) the quality by which each service is provided. QoS is described through the selection of a set of QoS parameters, specification of QoS target values and the choice of QoS measurements and evaluation mechanisms. A QoS parameter is a variable that characterises QoS.

In a multi-provision environment, the QoS provided by one entity may depend on adequate operation of other entities. Thus, there may be a need for relating the QoS levels obtained for the different interfaces, as shown in Figure 4. The QoS delivered at the user-primary provider interface *(QoS^u)* depends on the QoS delivered to the primary provider by its sub-provider(s) *(QoS^s)* as well as the influence of the QoS mechanisms locally implemented in the primary provider's domain *(L_QoS)*. Naturally, multiple interfaces may exist between one pair of actors in a configuration. However, due to the potential dependencies between interfaces, mechanisms for supporting the agreed QoS at the interface may be implemented in a number of ways.

Figure 4 Schematic illustration of QoS related to interfaces between (groups of) entities



It is emphasised that although dependencies are identified between QoS referring to different interfaces, a single provider is responsible for the aggregated QoS towards a certain user. This is the "one-stop responsibility" concept.

Naturally, a user may see various levels of details of services (and service components). Having a clear view of which service that is provided at an interface is essential. That is, the service provided has to be unambiguously specified. The QoS has to be described accordingly in the QoS-related part of interconnection agreements, named QoS Agreement (QoSA). In principle, the QoSA should contain the information on all the business/technical interfaces, traffic patterns, QoS parameters with objectives, measurement schemes, and reaction patterns, as depicted in Figure 5.

Interfaces description includes the description of all the interaction points relevant for the agreement – both business and technical. It might contain the information on the service delivery point, protocol(s) to be used, measurement points, observation points, points where a reaction pattern will be applied, etc.

Traffic pattern describes the characteristics of the expected traffic flows. This information allows the provider to manage resources in its domain in order to deliver the service with the appropriate QoS. The description of the traffic should envelop both application and management information flows. Traffic patterns can be described on different time scales (e.g. during the day, per service instance, etc.).

The description of **QoS parameters** and objectives implies expressing the performance of a service by assigning values to a number of QoS parameters [4]. Three categories of QoS parameters are identified, ref. [5]:

- Speed characterises the temporal aspects.
- Accuracy characterises the degree of correctness with which a given function is realised.
- Dependability characterises the degree of certainty that a function is performed.

Considering QoS objectives, they can be specified by target values (e.g. total maximum delay), or by thresholds set to a QoS parameter (e.g. an upper/lower bound), and so forth. The QoS objectives may be expressed as strict guarantees

or in a looser manner. Since QoS objectives are closely related to both measurements and reaction patterns, measurement procedures and conformance rules should (e.g. statistically) fit the granularity set to the QoS objective.

The **measurements** description may include: the identification of relevant measurement points, the specification of the measurement environment, description of the technique(s) for obtaining the measured values, specification of the methodology to present and evaluate the results by parameters, and the method to be used for taking decisions on acceptance based on the level of compliance of the measurement results with the stated requirements and commitments.

A set of **reaction patterns**, related to failure to meet either traffic patterns or one/more of the QoS objectives should be described. Such a description may include the reaction patterns both for cases of detecting the non-conformant traffic and detecting the service degradation. The reaction patterns for both entities should be stated including the inputs to initiate the reaction (e.g. results of measurements), related constraints (the duration, timeliness, type of actions), and the description of the reaction itself. The reactions could be technical (policing the traffic flows, suspending or aborting the activity, sending alarms, warnings, etc.), economical (e.g. discounts, initiation of using compensation schemes), legal and ethical (e.g. publishing the “antispam black lists”), etc.

One may observe that the QoS-related part of an agreement expresses how the user is supposed to behave, how the provider is supposed to behave, how to measure their behaviour and the compliance with the agreed requirements, and how to react in case the behaviour of any of them is not according to the conditions given in the agreement. Examples of such agreements are given in [2].

3 Relations with International Bodies

As already stated, there are many approaches to QoS that are considered by organisations and standards bodies. The aim when creating the EQoS framework was to develop a framework that could harmonise current positions wherever possible. Forums involved in QoS issues within which projects and studies have been investigated and which have liased with EQoS are: the ATM-Forum, COST, DAVIC, ETNO, ETSI (TIPHON, NA4, STQ), IETF, EURESCOM, INTUG, ISO (ODP, JTC1), ITU (SG13, SG2, QSDG), TMF, OMG, TINA, EU, National Regulatory bodies.

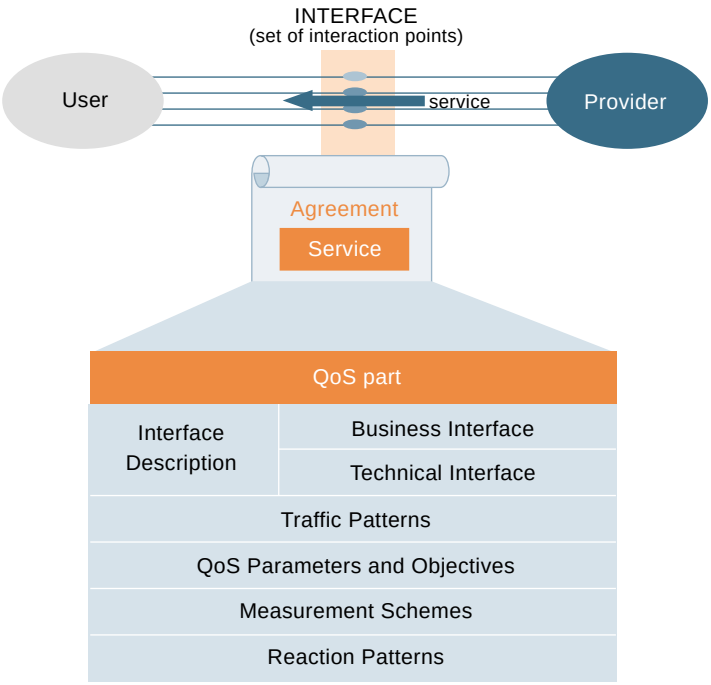


Figure 5 Agreement between user and provider, particular emphasis on QoS-related aspects

Some international bodies that have made efforts towards making QoS frameworks are:

- ITU (International Telecommunication Union) where part of Rec. E.800, [6], is adopted by IEC (International Electrotechnical Commission) as terminology standard IEC 191. Relevant results are also found in Rec. X.641.
- ETSI (European Telecommunication Standards Institute) framework, [4]. This framework is e.g. based on the work of the FITCE (Federation of Telecommunication Engineers of the European Community) Study Commission.
- The ISO (International Standards Organisation)/OSI (Open System Interconnection) QoS framework [7].
- The Telecommunication Information Networking Architecture Consortium (TINA-C) QoS framework [8].
- ETNO (European Public Telecommunications Network Operators Association) Working Group 07/95 on QoS, which is working on a consistent set of QoS parameters. The aim is to harmonise European QoS definitions and possibly performance targets for pan-European services, in order to facilitate comparison of the results of the measurements. The work is based on the approach of the FITCE Study Commission and ETSI. The work has hereto concentrated upon voice telephony.

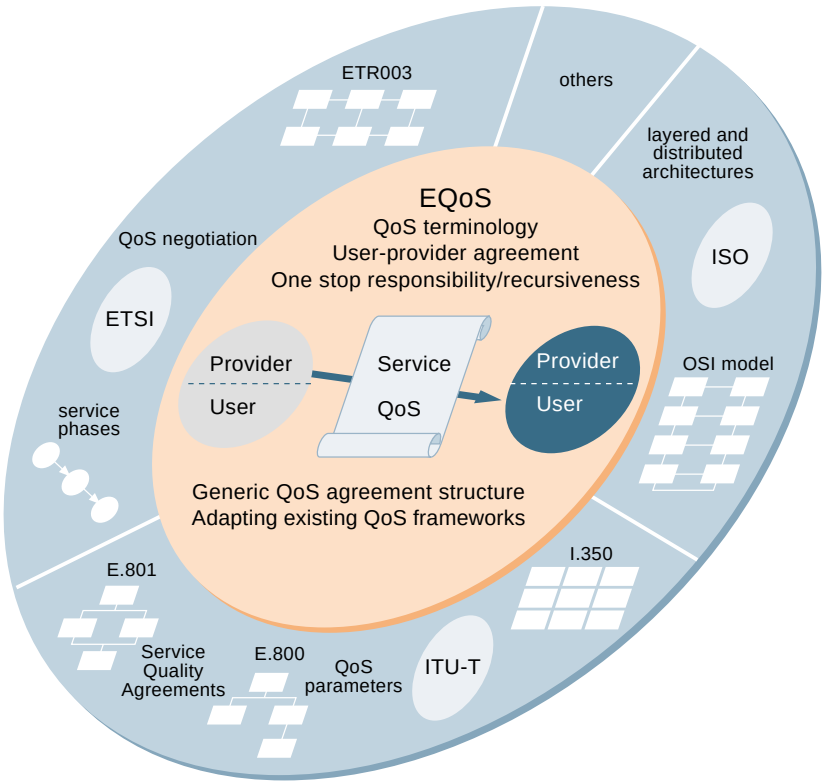


Figure 6 The EQoS framework relates to other QoS frameworks

- EURESCOM (European Institute for Research and Strategic Studies in Telecommunications), which has addressed QoS issues in a series of projects (see references in [1]).

It is to be noted that no QoS framework from the Internet Engineering Task Force (IETF) has been identified. Although a number of documents is dealing with QoS issues, they are not organised in an agreed IETF QoS framework.

An illustration of relationships between the EQoS framework and diverse contributions from ITU-T, ISO and ETSI is depicted in Figure 6.

From ITU-T Rec. E.800, the set of QoS parameters can be seen to be applicable in a general context, i.e. for provision of any telecommunication service. Details about organising parameters are found in other recommendations, e.g. ITU-T Rec. I.350. The notion of various aspects of QoS parameters and activities is found in ITU-T Rec. X.641. The main content of QoS agreements is based on ITU-T Rec. E.801.

The OSI model as presented by ISO, gives a presentation of a functionally layered and distributed architecture. This is widely applied as a conceptual model for communication between elements in a network. In addition, relationships between different functional levels are given. This is an origin of the recursive application of QoS agreements as defined in the EQoS framework.

Relationships between the various viewpoints of QoS can be found in ETSI ETR 003, presenting the impact from several actors (service provider and network operator) on the QoS. In addition, relationships from required/offered/achieved/perceived QoS are given. That is, the QoS negotiation process can be identified (prior and during service provision). Relating QoS to various phases in the service life cycle is also treated in ETR 003.

4 Further Issues

Having a generic QoS framework presents a basis for discussing and solving the open issues. A number of issues that request further effort is addressed in this chapter. No priority is given to either issue, implying that none of them should be considered as of higher importance compared to others.

4.1 Business Matters

As described in Chapter 2, deciding upon QoS should be included in the business activities. Moreover, when settling the appropriate values of QoS parameters, one would likely face several trade-offs, like what QoS level to offer at what price, which QoS-related mechanisms to implement within its domain, and so forth.

Carrying out business-related evaluations, the market situation is taken into account including potential customers' requests, competitors' activities, regulatory directives, etc. Being aware of that service degradations/failures may occur, stating conditions in the agreements can be looked upon as risk considerations. That is, damages/penalties in case an event happens are balanced against the cost of the means undertaken in order to lower the probability of the event occurring. Lower cost is commonly sought, while major negative consequences are avoided. Balancing internal mechanisms and the conditions stated in agreements towards any sub-providers would also be part of this picture as seen by a provider. Which mechanisms to activate within a provider's domain (and accompanying cost) can then be traded off against the costs of the service components (and QoS) that sub-providers deliver. This challenge may be considered an optimisation problem, possibly also including time/evolution.

The objective function to be optimised may indicate the profit obtained by the provider. The set of variables would reflect the options that the provider may face, both the possibility of getting service components from sub-providers with corresponding QoS, and the internal mechanisms that can be activated and their corresponding influence on QoS (ref. expression in Figure 4). Schematically, the objective function should then be maximised given a set of constraints,

like which sub-providers are available, which mechanisms are possible, how service components can make up the service, etc.

In addition, an optimisation problem could consider certain instances of time. Levels of uncertainty would likely be attached to the different instances as there may be a number of further evolution paths from an instance and each path could have a likely profit level. As the number of combinations facing a provider grows, appropriate abstractions and procedures are needed in order to solve the optimisation problem.

A service may be specified in a variety of ways. Increasing the granularity of service specification restricts the degrees of freedom left for the provider to establish the service. However, the granularity of QoS involved also increases. In other words, a high-level service specification requires a corresponding high level of QoS parameters. A challenge would be how to identify the granularity level in correspondence with the service specification. In order to encourage re-use, similar generic structures regarding all services could be asked for.

4.2 Choosing QoS Level Based on Consequence/Risk

Users do not commonly specify their quality needs merely from a communication point of view, but rather from the consequences they envisage on their business and human relations – the secondary scope – if a failure occurs. This also determines the level of QoS they are willing to pay for regarding a specific service ordered.

Then, guidelines to a way of assisting the customer in structuring the consequences they might envisage, both in primary and secondary scope, would be requested. The generic components of a service failure are elaborated in [9]. One such generic component may have greater impact on a customer perception of consequence than another.

To perform a dependability evaluation of telecommunication systems, insight is needed into phenomena causing outages (e.g. error propagation and nearly simultaneous trigging of faults in software). The strong logical interdependencies in some systems make them vulnerable to these kinds of failures.

End-user service failures may be caused by some failures within the network, each of which can be caused by a range of faults. In addition, the fault tracing will be dependent on a number of factors like: network topology, O&M routines, traffic situation, etc. although demanded by all service providers, a general root-cause analysis (RCA) of service failures still awaits further research.

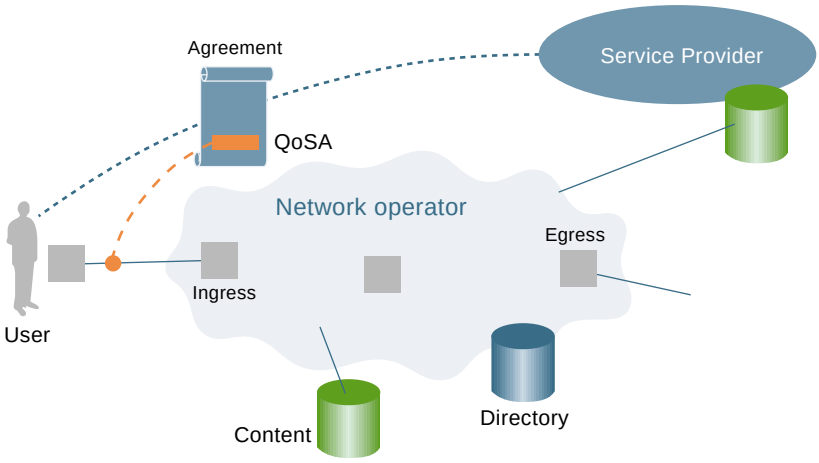


Figure 7 Several functions/ components could be involved when realising a service sold to an end-user

4.3 QoS Parameters that Communicate with the Customer

QoS parameters that communicate well and unambiguously to the non-professional users – i.e. residential end-users, are of particular interest. Imperative to all parameters is that they are harmonised and allow for mapping into others. Monitoring a QoS parameter could be done both "continuously" and by "sampling" according to "typical" usage.

The dynamics of the QoS may have a large influence on how the QoS is perceived. A QoS parameter could be described in several ways of statistics, like average(s) over some intervals, variance, instantaneous distribution of the parameter (occurrence frequency), and auto-correlation (the auto-correlation indicates how rapidly the QoS level changes and may give valuable information on how the QoS is perceived).

With respect to the concept of reaction pattern special emphasis should be put on measurements of the extreme behaviour of the QoS process – or more specifically – when the values of the QoS parameters approaches (or are in) a range that represents a service failure.

An essential question is which service is really sold. In several cases, pure transport of IP-packets is not the service seen by an end-user. That is, a "higher" level service, possibly aggregating several "IP flows", could contribute to the service as experienced by a user. This is illustrated in Figure 7, where the end-user service is agreed between an end-user and a service provider. But, the end-user service includes not only IP-packets transport service provided by a network operator, but also the informational service provided by the service provider, e.g. content.

The QoS parameters would then have to be specified on a corresponding level and specified in

the QoS part of an SLA, i.e. QoSA. This may ask for a number of mechanisms/features that should be available in the network elements and management systems. Taking into account that several traffic flows (possible to different destinations) could be involved, collecting the relevant parameters would be a non-trivial task.

4.4 Aggregated QoS Parameters

As seen from ITU-T Rec. E.800 the upper part of the "parameter hierarchy" represents parameters that are supposed to communicate with the human users. When rising upward in the hierarchy those parameters get an aggregated nature. The aggregation is so far not taken care of properly by any standardisation body. At the moment QoS is defined by a chosen set of parameters – not all of them having the same dimension.

To allow for a simple yet unambiguous parameter communication to the end user an aggregation procedure is needed to allow for a number of pertinent parameters to be combined into aggregated parameters. One way of aggregating the parameters is to introduce weights. This weight should correspond with the importance this parameter has to the customer. However, in general, such a mapping could be rather complicated. Such a formula calls for the same dimension on all parameters. A possible solution could be to reformulate the parameters into a dimensionless fraction or index. A specific advantage is that such a formulation makes the quality assurance aspect of the measurements stand out clearly.

4.5 Traffic Patterns that Apply for QoS Agreements

Traffic involved in providing a service to a user is specified in the agreement. Examples of characterising traffics and their aspects are given in [2].

In the agreement QoS guarantees could apply for traffic compliant with the requests of the user. However, the guarantees do not apply in case the user utilises the service outside its scope. That is, there will always be restrictions on both provider and user for QoS guarantees to be effective. Therefore specification of the traffic pattern may be done in a way that closely reflects the customer's needs in normal situations, emergency/extraordinary situations, and, service failure situations (i.e. when QoS is below agreed level). Also the user has to react to her/his provider with the agreed feedback information paramount to keeping the agreed QoS level. The feedback information might represent services and will have to be included in the agreements as such ones.

4.6 Agreements Referring to Common Resource

Another interesting issue opens when several independent agreements refer to the same resource group. Considering the configuration depicted in Figure 8, two service providers may have made agreements with an end-user. Then, both services are to be provided via the same access line. Utilising each service sequentially may work well. However, in case both services are used simultaneously the capacity of the access line may not be sufficiently high to keep the guaranteed values. Then, some means for managing such situations would be requested.

Naturally, the end-user could detect and be able to sort out the situation. In addition, mechanisms within the network are sought both in order to support the end-user and to be able to document the events taking place. On one hand, this may be though of as being able to map from the parameters as observed in the network to the parameters described in the agreements. Taking into account flexible services, like IP-based, that kind of mapping may not be simple to devise. The potential ambiguous mapping is illustrated in Figure 8. Mapping from the conditions stated in the QoSAs to network performance parameters (PPns) could be uniquely described. For example, delay between two points as described in QoSAs could be followed by a corresponding parameter as monitored by the network. On the other hand, mapping from the set of performance parameters, PPns, to the set of QoSAs may not be obvious, as there may not be any indication of which agreement a certain IP packet belongs to.

The multi-provider configurations that involve a common set of resources advocate the need for introducing additional features for solving the issues described above. A particular gain would be obtained by a network operator incorporating efficient means for dealing with these aspects.

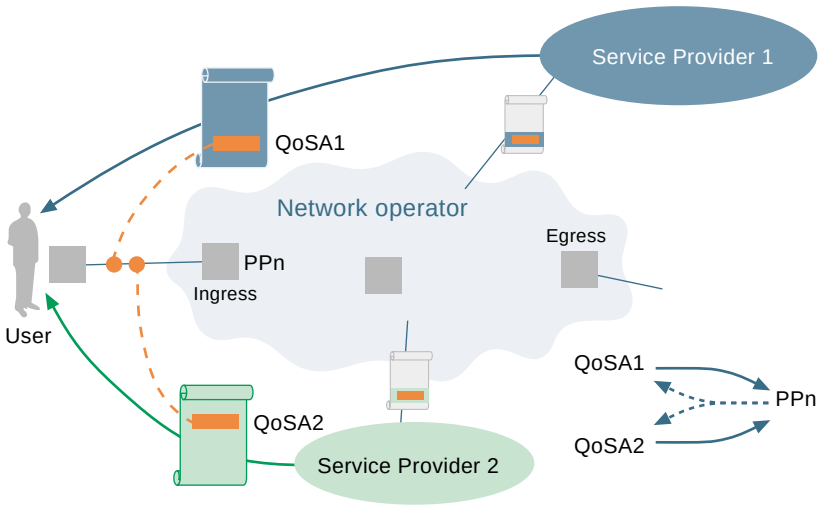


Figure 8 Several agreements referring to a common set of resources

5 Concluding Remarks

Managing QoS is a growing challenge in environments where ever-growing number and types of providers/users/services/applications is present. A procedure for deciding on QoS is outlined in this paper, along with the description of the EQoS framework. This framework establishes not only the terminology and principles (e.g. one-stop responsibility, recursiveness) but also incorporates major frameworks and QoS related work published by different groups, bodies and fora. It is generic in a sense that it is applicable on all types of services, actors (playing roles of both users and providers), and technologies involved in the complex multi-provider configurations. The applicability of the framework has been demonstrated on several examples within the EURESCOM P806-GI work [2]. The EQoS framework has been promoted in various international arrangements. Since a rationale behind the framework is to align the telecommunication business with other business areas, it is believed that it provides results pointing in the proper direction for the further QoS work. A number of QoS-related issues ask for further investigations, and some of those are discussed in this paper.

References

1 EURESCOM. *EQoS – A common framework for QoS/NP in a multi-provider environment.* Heidelberg, September 1999. (EURESCOM P806-GI Deliverable 1.)

2 EURESCOM. *How to apply EQoS? Case study VoIP. QoS handbook.* Heidelberg, September 1999. (EURESCOM P806-GI Deliverable 4.)

3 ITU-T. *Framework for service quality agreement,* 10/96. (ITU-T E.801.)

4 ETSI TC-NA. *Network Aspects : general aspects of Quality of Service and Network Performance.* October 1994. (ETSI Technical Report ETR 003 (ref. RTR/NA-042102).)

5 ITU-T. *General Aspects of Quality of Service and Network Performance in Digital Networks, including ISDN,* 03/93. (ITU-T I.350.)

6 ITU-T. *Terms and definitions related quality of service and network performance including dependability,.* 08/94. (ITU-T E.800.)

7 ISO. *Quality of Service Framework.* (ISO IEC JTC1-SC21.)

8 TINA-C. *Quality of Service framework.* (Doc. no TR_MRK.001_1.0_94.)

9 Espvik, O et al. *User perception and service failure – QoS aspects up to 1998.* Kjeller, Telenor R&D, 1999. (R&D note N 60/99.)

Some Quality and Coverage Problems in Audio Broadcasting

KNUT N. STOKKE

After graduating from the Norwegian Technical University (Trondheim) in 1958, Knut N. Stokke (70) worked from 1959 to 1969 in the Planning Division of the Broadcasting Office of the Norwegian Telecom Administration, and thereafter with the Transmission Section where his activities included specifications and regulations for broadcasting transmitters and transposers. In 1987 he joined the new regulatory organisation, the Norwegian Telecommunications Regulatory Authority, where he was head of the Section for Broadcasting. Knut Stokke has been a member of the Norwegian delegation to the major broadcasting conferences, and has also participated in various ex-CCIR Study Groups and more specifically Study Groups 5 and 6 (now St.Gr. 3). Knut Stokke retired 1 March 1999.

Introduction

When we are planning broadcasting coverage in an area, we normally begin by examining the transmission conditions. The available bandwidths are given in international agreements, and we also refer to the other parameters given there. However, are there other conditions that may influence the strength and the quality of the received signal?

In order to simplify the considerations, we have in Figure 1 indicated how the coverage may be from a transmitting station placed in an area with homogeneous conditions in all directions. The coverage is here divided into three areas: the near coverage area, an intermediate coverage area, and a distant coverage area. Is it possible that we may have contradictory interests between these areas concerning broadcasting quality and broadcasting coverage?

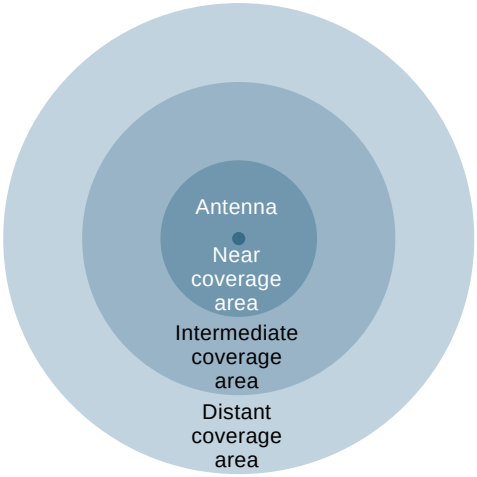


Figure 1 Coverage when the transmission conditions are the same in all directions

Special Conditions in AM Broadcasting

In 1975 there were some modulation tests at the Oslo AM transmitter station. At that time the frequency was 218 kHz and the radiated power was 200 kW e.m.r.p. (effective monopole radiated power). These tests were carried out by a group consisting of members of the Norwegian Broadcasting Corporation and the Norwegian Telecommunications Administration.

The reason for performing the tests was that there were complaints concerning the audibility of the transmissions. One of the plaintiffs said: We hear the carrier but not the modulation.

The first investigations showed that the mean modulation for the Oslo AM transmitter was only about 30 %. It was proposed to use a *limiter* (see Figure 2) for the modulating signal. This would cause a reduction in the dynamic range, but at the same time the mean modulation would increase causing an increase in the sideband power and thereby better audibility. *As long as the carrier is not severely overmodulated, it is*

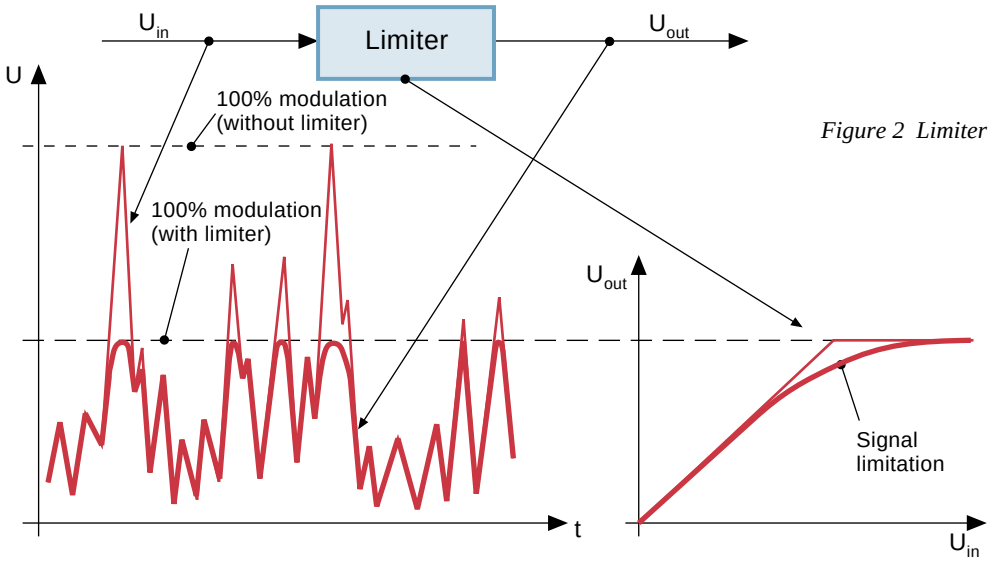


Figure 2 Limiter

the sideband power which determines the demodulated signal strength and consequently the audibility.

It should also be mentioned here that more advanced limiters are used today. For instance some of them have the possibility to limit the different frequencies independently.

There were some reluctance concerning the use of limiters, just due to the reduction in the dynamic range of the signal. It was therefore decided to use relatively moderate limitation. However, the mean modulation was raised from about 30 % to about 70 %; that is, the sideband power was increased by 6 – 7 dB, and the test should then compare the receiving conditions with and without limiter at the transmitter.

Because we wanted professional people to judge those tests, the Radio Interference Division in the Norwegian Telecommunications Administration was engaged to listen to the transmissions. In addition they had people at different locations in the coverage area of the Oslo AM transmitting station. In order to avoid biased judgements they were given as few details as possible. They were told to use field strength measuring equipment, a normal AM receiver, and a variable attenuator to judge any subjective variations. They were also instructed to describe as exactly as possible all that happened to the AM signal at the changes.

The time interval for the changes between transmission with and without limiter was 2 minutes. In order to avoid any doubt about the time for changes, the clocks were synchronised.

The tests were done in the daytime, that is, it was mostly the ground-wave which was investigated. Then it is the natural electromagnetic noise which limits the signal-to-noise ratio. However, the use of a limiter would be even more advantageous for transmission via the ionosphere where interference normally determines the audibility.

We got some unexpected results. Most of the listeners said that the dynamics of the signal seemed to increase when the signal (that is, the sideband signal) strength increased. In fact, we then reduced the dynamic range of the signal. We had to give this result some thought before we got into the explanation of the problems, and the answer is that we have to consider the coverage range for the signal and the quality of the signal.

If we look at a field strength curve, for example the thick solid curve in Figure 3, this curve is *a mean curve*, that is, a curve measured with a slow instrument. If we had an instrument that could follow the modulation, this instrument would swing between the absolute values of the modulation as also indicated in Figure 3. These peak values are relatively the same for different distances. We may here observe that we need not move so far away from the transmitting station before a rather large part of the dynamic range is coming into the noise, and this part of the dynamic range is of no use.

If we now increase the transmitter power by 6 – 7 dB, the mean curve will increase with the same value, for instance to the thick dashed line in Figure 3. We then observe that at the same time more of the dynamic range is over the noise level, giving more usable dynamics to the listeners.

It does not matter whether the increase in the radiated power is due to the increase of transmitter power or to the increase in sideband power, on the assumption that we have no severe overmodulation.

Another unexpected result was the subjective assessment of the increase in the signal strength. Most of the listeners thought that the increase in signal strength was about 6 – 7 dB, but in about 130 km from the Oslo AM transmitter the observers thought that the increase was as high as 10 – 15 dB. Because several observers at about the same distance had the same impression, it was necessary to find the reason for this phenomenon.

The AM group organised these investigations. We began measuring near the transmitting station and were a bit surprised that the 6 – 7 dB could be observed subjectively also near the

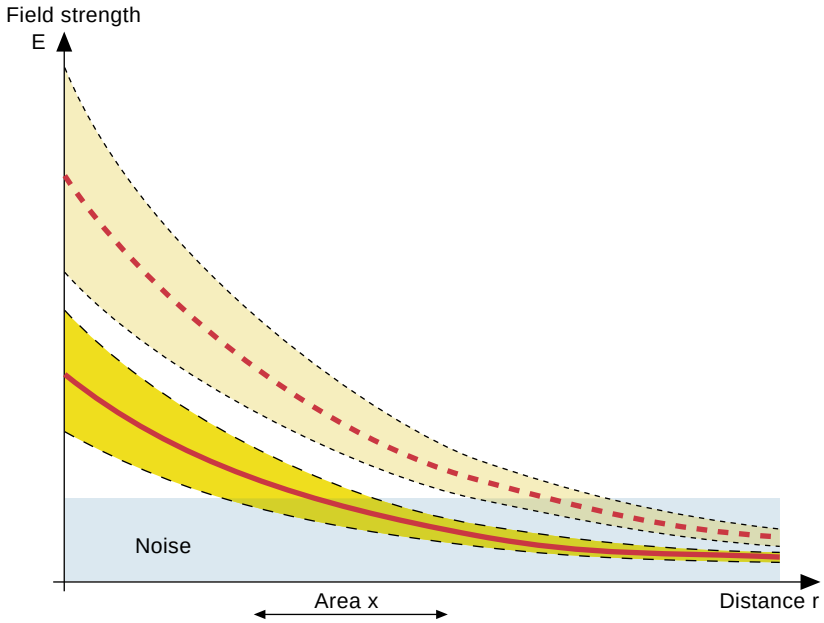


Figure 3 Field strength curves

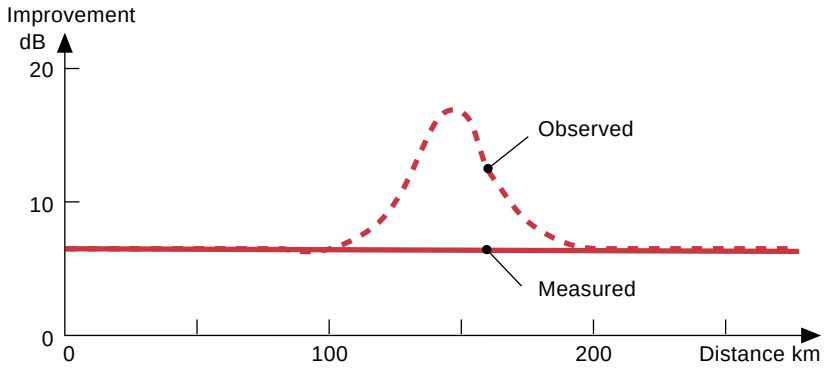


Figure 4 Subjective and objective measurements

station. We had not expected 6 – 7 dB to have any noticeable effect in the large field strengths near the transmitter. An explanation may be that the noise in this frequency band is higher than indicated in Figure 3.

The improvement both for subjective and objective measurements was about 6 – 7 dB until we were about 100 km from Oslo. Then the observed signal strength seemed to increase from the earlier 6 – 7 dB, and in some areas at about 150 km we observed more than 18 dB. This large difference between subjective and objective measurements may be explained by the fact that in this area (area x in Figure 3) the signal was increased so much that the signal-to-noise ratio was changed from not usable to usable.

Several other conditions influence the audibility of an AM broadcasting station. Because of small modulating signal bandwidth (4.5 kHz to the –3 dB point for broadcasting in the long and medium waves, and 5 kHz to the –3 dB point in the short-wave band) the form of the amplitude curve for the modulating signal is of great importance. Without shaping, that is, with linear amplitude curve for the modulating signal, we get too much bass and too little treble. Using a moderate shaping which here means extra amplification of the highest modulating frequencies, will give a better balance between bass and treble and consequently better audibility.

With respect to quality and coverage, we will here have conflicting interests. *If we look at Figure 1, only the near coverage area will have full advantage of high quality transmissions*. Further away from the transmitting antenna, in the intermediate and the distant coverage areas, the signal-to-noise ratio will be reduced to values too low for high quality transmissions. Here we have to decide which of the areas are the most important for the transmissions. It may not be appropriate to go for very high quality which is of advantage only for listeners near the transmitting station. It may be more advantageous to reduce the quality requirements in order to give the intermediate and the distant coverage areas a reasonable signal strength and quality.

There should be a certain liberty to choose between quality and coverage, and this must be discussed when planning the transmissions. However, it should be pointed out that when a broadcaster will go for high quality transmissions, the values must not be so extreme that it will lead to unacceptable use of the frequency resources. It will then be a matter for the frequency regulation authorities.

High quality transmissions in low frequency, medium frequency, and high frequency broadcasting, is possible by using digital coding and compression. If we at the same time introduce stereo, there will be very little difference between such transmissions and high quality FM transmissions, especially when listening in cars.

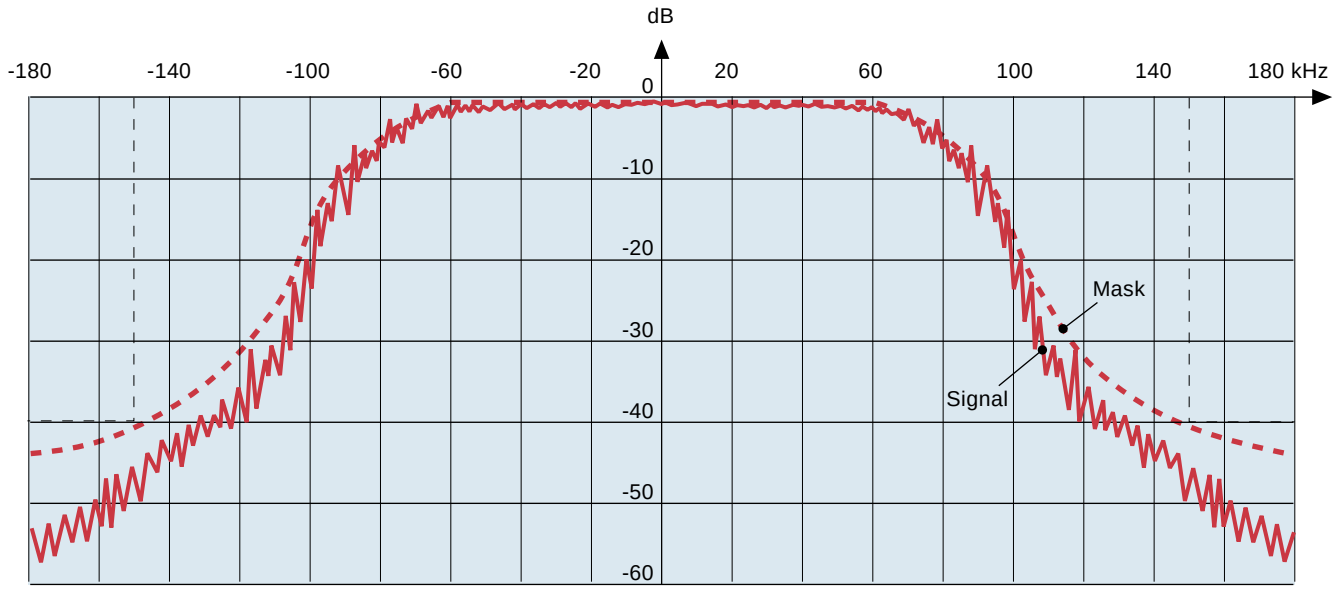
Special Conditions in FM Broadcasting

In FM broadcasting we have about the same problems as for AM concerning coverage range and received signal quality. However, because FM is a relatively complicated modulation system, it is rather difficult to get a clear view of the modulating process.

For instance the signal from an FM transmitter is dependent on both the amplitude and the frequency of the modulating signal. Therefore it is difficult to measure the deviation with a simple instrument, for example a deviation measuring equipment. We may have good results when such an instrument is used directly at the FM transmitter, but if we have other signals near the measured channel, a deviation measuring equipment will also take these signals into account. In principle a deviation measuring equipment has to be rather broad-banded because it shall measure frequency deviation.

The measurements of the deviation of the FM sound signal in a television signal (with 50 kHz deviation) have normally been done with a deviation measuring equipment. However, it is then possible that we may get the luminance signal, the colour signal, the NICAM signal, etc. into our measurements. This may be the reason why many television transmitters have rather low deviation for FM sound signal. It is here more suitable to use a spectrum analyser to see the real deviation. A limiter for the sound signal may also be advantageous, not only for the modulation, but also because we get less interference to the colour and the luminance signals.

In Figure 5 is shown the frequency envelope curve for a well modulated FM broadcasting transmitter (with 75 kHz deviation). A limiting



curve, a mask, is also given, and the signal curve should be kept within these limits. However, it is not so easy to consider the envelope curve for such an integrated signal. As mentioned before, the deviation is dependent on both amplitude and frequency for the modulating signal, and there will often be a discussion if the mean value of the signal or the peak values should be used. If we go for the latter, we have to be aware of the fact that the results will very much depend on the accuracy of the instruments.

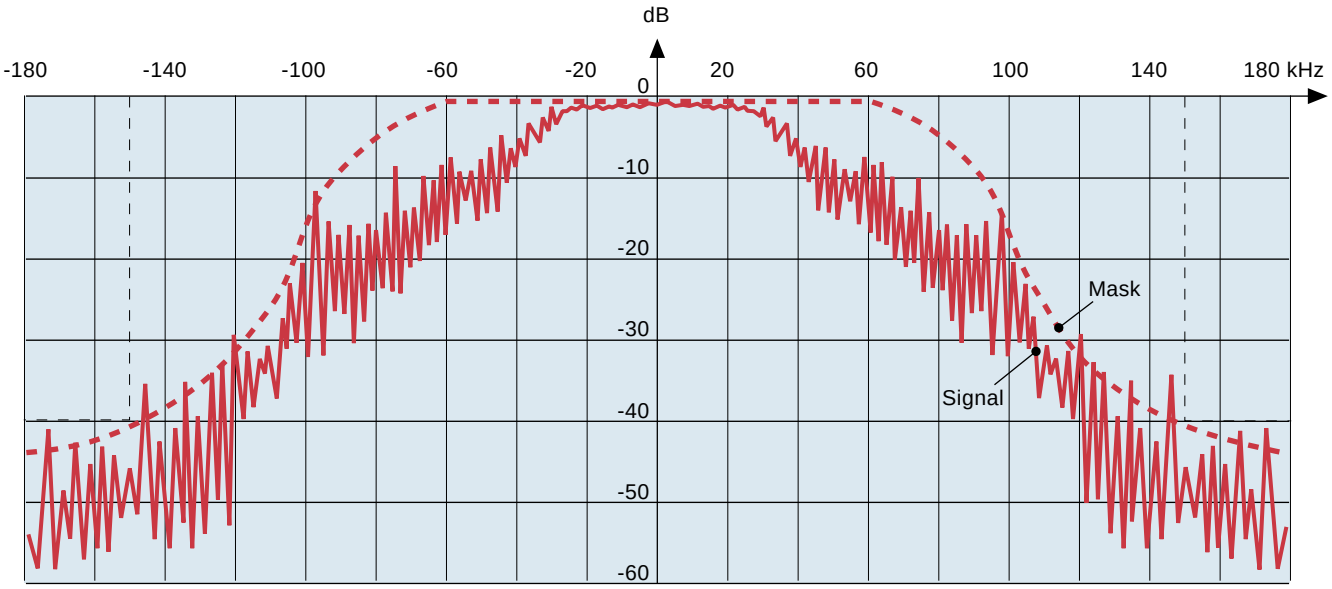
It is not of great importance for the signal strength or for the interference conditions if the signal exceeds the mask within the actual channel. The most important is that the specified values are fulfilled in the adjacent channels giving low interference to other transmissions. This necessitates a limiter in the modulating signal. *This limiter begins to influence the signal at ±100 kHz, that is, outside the necessary bandwidth for the FM signal, 180 kHz.*

Even a small limitation will give a rather large increase in audibility. Therefore it may be more favourable to use a little limitation in the modulating signal instead of increasing the bandwidth some kHz.

In Figure 6 is shown how the signal may be from an FM transmitter without limiter (free dynamic conditions). We see here that the usable deviation has to be very much reduced in order to keep the signal within acceptable limits. In this example the usable deviation is only about 30 kHz. *And because the FM demodulators in the receivers give a received signal strength which is directly proportional to the deviation, the audibility will be extensively reduced in comparison to the use of limiter.*

Figure 5 Mask and envelope curve for a well modulated FM broadcasting transmitter

Figure 6 Mask and signal curve for an FM transmitter without dynamic limitation



These circumstances may lead to odd results. A well modulated small FM transmitter may be heard with a larger signal strength in the receivers than a high power FM transmitter with low modulation level, provided that the field strength is sufficient to give acceptable reception. In one case an FM transmitter with 10 W effective radiated power (e.r.p.) gave better audibility than an FM transmitter with low modulation level but 100 kW e.r.p.

If we want all the FM transmitters to be received with about the same audio level in the receivers, all of them should have the same deviation conditions. But again we have to consider the audibility and the quality of the signal. If we want to sacrifice a little of the quality in order to get better audibility, this is possible on the condition that the signal is kept within given limits.

If we want to increase the quality of an FM signal by reducing the deviation, we will get poor utilisation of the channel. Looking at Figure 1, it is only the near coverage area that will be an advantage of such transmissions. The intermediate area and the distant area will get lower signal quality because of too low signal-to-noise ratio. It is again necessary to consider quality and range of the signal.

In most European countries it has been agreed that the necessary bandwidth of an FM broadcasting transmitter is (Carson's formula):

$$B = 2(d + f_m) = 2(75\ \text{kHz} + 15\ \text{kHz}) = 180\ \text{kHz}$$

where B is the necessary bandwidth, the deviation parameter d is 75 kHz (in the rest of Europe 50 kHz), and the highest transmitted modulation frequency is f_m = 15 kHz. This means that we take an extra component outside the deviation parameter 75 kHz into account. To consider higher components in the FM signal will give very little influence on the signal quality. These components are relatively small, and in addition the signal-to-noise ratio for these components will be very low.

It may be mentioned that for planning purposes an FM broadcasting channel bandwidth was agreed to be 300 kHz.

Comments Concerning DAB (Digital Audio Broadcasting)

If we want to have high quality audio broadcasting, digital audio broadcasting is a possible solution. Up to now there is no other new system which may compete with DAB. It is also easy to regenerate digital signals, and we have full quality for the received signal until the field strength is so low that the transmission is interrupted. And because the retarded signals or reflections are used in a positive way, the reception of DAB in mountainous and hilly terrain is totally superior to the reception of FM broadcasting. This is especially observed in mobile reception. In flat terrain without reflecting objects the difference is rather small.

It may also be mentioned that when it was agreed to introduce DAB, this should be done within the already existing frequency band, 87.5 MHz to 108 MHz. Some countries had already begun to allocate frequencies for DAB in that band. However, the BBC in Britain had done some tests and measurements for DAB in different frequency bands, and had come to the conclusion that frequencies in the 100 MHz band were not especially advantageous for DAB. The results were better at higher frequencies, and it was necessary to find a higher frequency band that could be allocated to DAB. *TV channel 12, 223 MHz – 230 MHz, and the band 230 MHz – 240 MHz seemed to fulfil the requirements.*

Conclusion

When planning and operating audio broadcasting, it is important to consider very carefully the quality and the coverage range of the signal. A reasonable balance between them will give a good utilisation of the frequency resources.

References

Kennedy, G. *Electronic communication systems.* Tokyo, McGraw-Hill Kogakusha, 1970.

Stokke, K N. *Senderteknikk.* Sandvika, Vett & Viten, 68–70, 97–103, 1989.

Stokke, K N. Måling av deviasjon og bandbredde for en FM-nærradiosender. *Telektronikk*, 83 (1), 82–88, 1987.

Stokke, K N. A method of measuring deviation and bandwidth of FM broadcast transmitters. *EBU Technical Review*, 216, 1986.

Stokke, K N. Considerations concerning the efficiency of FM broadcasting transmissions. *ITU Telecommunication Journal*, 60 (VII-VIII), 1993.

Status

Converging Broadcasting and Telecom

PER HJALMAR LEHNE

Per Hjalmar Lehne (42) is Research Scientist at Telenor Research & Development, Kjeller. He is working in the field of personal communications, with a special interest in antennas and radio wave propagation for land mobile communication.

per-hjalmar.lehne@telenor.com

One of the trends of today is the convergence of broadcasting and telecom. While broadcasting is a one-way service with an identical service to all users, telecom is the opposite, namely individual services to individual users.

With the rising demand of bandwidth and multimedia services in the telecom world, broadcasting technologies have become more and more interesting as candidates for point-to-multipoint (multicast) services. However the one-way nature of broadcasting methods has been a major drawback.

In order to solve this, different approaches have been studied and proposed. Most proposals utilize another network or technology (e.g. PSTN, ISDN, GSM, ...) to provide the so-called return channel.

Several projects within the European Framework Programmes have been completed and new ones are ongoing. In the ACTS programme, the *MEMO* project has established an architecture and APIs for a combined DAB/GSM network. Trials were performed in the 1996 to 1998 period. MEMO realised *iDMB* using a GSM 9.6 kb/s circuit switched interaction channel and 1 Mb/s downstream via DAB. DAB can potentially supply up to 1.7 Mb/s. The MEMO project also developed the terminals and software needed. The focus of the MEMO project was on mobility, the reason why GSM was chosen as the return channel technology. Three main service categories were studied: *Data broadcasting with virtual interaction, Interactive broadcasting,* and *Personal services*. For Personal services, the TCP/IP protocol suite End to End was used. The method of encapsulating IP packets in DAB was also proposed to Eureka 147 Forum. MEMO also did a prototype on combining DVB-T and GSM.

Another project in ACTS, *DIGISAT*, studied the implementation and use of a return channel for satellite for SMATV systems. Trials were performed in 1997 incorporating 600 users all over Europe demonstrating for example *Near Video on Demand* (NVoD) and *Co-direction*, the transmission of different views of the same event to the home, so that individuals can choose the desired view. The project worked directly towards the DVB project (see below).

In the fifth framework programme, IST, studies and research continue. The *NexTV* project ("New media consumption in Extended interactive TeleVision environment") looks at how media will be consumed in interactive digital television. It also aims at producing technical specifications of an end-to-end system. Another project is *myTV* ("Personalised services for digital television"). This project looks at services and technologies for consumer platforms with built-in local storage. One example is the ability to turn local storage into a personalised television channel with access to contents and services at the consumer's convenience, independent of the moment of broadcasting. Both these project will end in 2001.

Important work is going on in order to standardize integrated return channels. One of these initiatives is in the DVB project, which was initiated by EBU in 1993. In this issue of *Telektronikk*'s Status section, Vendela Paxal describes the work in the DVB-RCS group, as well as the reference model and architecture of the concept in the paper *"DVB with return channel via satellite"*. This concept describes a return or *interaction* channel via satellite providing from 64 kb/s up to 2 Mb/s data rate for *uplink*, thus providing a bandwidth at least as good as the aforementioned hybrid methods. Parts of the paper have previously been published by DVB as a 'White Paper' on RCS (DVB-RCS200).

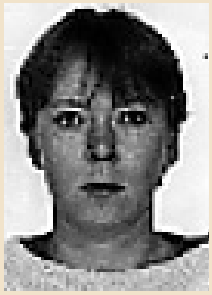
The DVB project also works on other return channel concepts, e.g. via PSTN/ISDN and GSM.

See following page for an explanation of the abbreviations.

ACTS	Advanced Communications Technologies and Systems (EU 4th Framework Programme) http://www.uk.infowin.org/ACTS/
DAB	Digital Audio Broadcasting http://www.worlddab.org/
DIGISAT	Advanced Digital Satellite Broadcasting and Interactive Services
DVB	Digital Video Broadcasting http://www.dvb.org/
DVB-RCS	DVB – Return Channel via Satellite
DVB-T	Terrestrial DVB
EBU	European Broadcasting Union http://www.ebu.ch/
GSM	Global System for Mobile Communications
iDMB	interactive Digital Multimedia Broadcast
ISDN	Integrated Services Digital Network
IST	Information Society Technologies (EU 5th Framework Programme) http://www.cordis.lu/ist/home.html
LMDS	Local Multipoint Distribution System
MEMO	Multimedia Environment for Mobiles
PSTN	Public Switched Telephone Network
SMATV	Satellite Master Antenna TV

DVB with Return Channel via Satellite

VENDELA PAXAL



Vendela Paxal (36) received her Maîtrise es Sciences degree in physics from the Université d'Aix-Marseille III, Marseille, in 1988; her Dipl.Eng. degree in telecommunications engineering from the Ecole Nationale Supérieure de Télécommunications, Paris, in 1990; and finally her Dr.Techn. degree from the Norwegian University of Science and Technology, Trondheim, in 1998. From 1990 to 1995 she worked with the digital signal processing group in SAT/SAGEM, Paris, where her main interests were advanced modulation and coding techniques. She designed ASICs for satellite and cable television receivers according to the DVB standards. In 1995 she joined the satellite and radio group at Telenor R&D, where she has been involved in various system studies, mostly within European projects. In addition, she has been working in the DVB Ad-hoc group for standardisation of the Return Channel via Satellite.

vendela.paxal@telenor.com

Introduction

Increased interactivity is a general tendency for telecommunication services today, an aspect which is also reflected in the more traditional distribution services like radio and television. Customers want to choose, sort, order, store and manipulate what they receive on their terminal, and ideally also interact from the same terminal. The distribution network becomes an asymmetric interactive network, with a possible evolution towards fully symmetric communication. This convergence between communication and broadcasting leads to an evolution from broadcasting to multicasting or point-to-multipoint communication, where the difference lies in the possibility to offer contents/services designed for individuals or groups of people with restricted access and billing. This evolution will also have consequences for satellite communications, certainly the most broadcast-oriented medium of all.

There are several ways to design a return channel for satellite multicast services, and many believe terrestrial return channels to be the most cost effective and practical. Commonly proposed terrestrial return channels are PSTN, ISDN or GSM. However, there is a large worldwide interest for a definition of a return channel via satellite, and there are several reasons for that. Firstly, as mentioned above, the "normal" consumer does not want to be bothered by technical set-ups with interconnections between TV, PC and telephone. A solution where all the technical equipment is concentrated within one box, and without having to fear blocked telephone lines etc. will certainly be appealing to many people. Another reason to choose satellite services is the increased traffic in the terrestrial networks, which often results in blocking or reduced quality of service. The instantly available capacity on a satellite link can, with efficient resource allocation, be set to 2 Mb/s for instance. A 100 Mbyte file will need about 7 minutes for transfer over satellite whereas the time required over a 64 kb/s terrestrial line will be about 3 1/2 hours. Finally, it is an advantage, both for the users and for the operators, that both channels are on the same medium. This enables better control with the QoS and the network management, the terrestrial infrastructure is often not controlled by the same operator as for satellite, and this is certainly not the case when national borders are crossed.

Standardisation Procedures for DVB-RCS

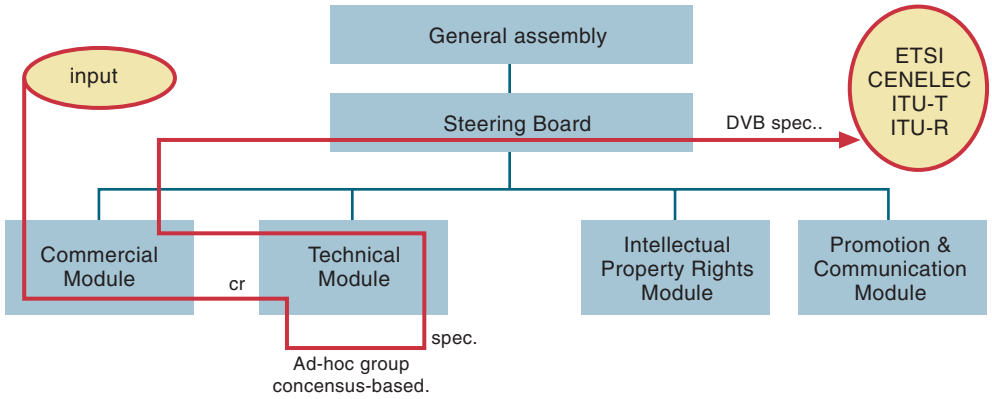
Due to the recognised need for a specification of a return channel via satellite, the DVB-TM (Digital Video Broadcasting – Technical Module) created an Ad-hoc group early 1999, called the DVB-RCS (DVB- Return Channel via Satellite). The DVB project itself was launched in 1993, under the auspices of the EBU (European Broadcasting Union) [1]. The motivation was to promote a common, standard European platform for digital TV broadcasting, and the idea was supported by all players; broadcasters, operators, standardisation bodies, media groups and the industry. Today, DVB counts 220 members from more than 30 countries around the world. Due to high professionalism, the DVB has gained confidence worldwide, even if initially intended for European purposes. DVB is not a standardisation forum itself – after approval in the DVB hierarchy, the specifications drafted in the various technical Ad-hoc groups are sent to bodies like ETSI or ITU to obtain the formal status as a standard.

Work on a standard is initiated within the CM (Commercial Module) of the DVB project before being considered by the TM, and actually many explain the success of the DVB project by this procedure. The TM will then create an Ad-hoc group, which will work for anything from months to years on the technical specification that will at best satisfy the CR (Commercial Requirements) issued by the CM. When the work in the Ad-hoc group is considered satisfactory by the group itself, the specification will be sent up through the DVB hierarchy for approval (see Figure 1). Both the CM and the TM may ask advice from the IPRM (Industrial Property Rights Module) on licensing and patents. After a final approval in the Steering Board, the draft will be issued for consideration within the appropriate standardisation organisation. The steering board is nominated by the general assembly, where all members are represented after having signed an MoU (Memorandum of Understanding) and paid the fee.

Contents of the DVB-RCS Specification

As for all other projects within the DVB, the work in DVB-RCS is based on the Commercial Requirements issued by CM. Three user profiles

Figure 1 The organisation of the DVB project and the typical life cycle of a specification

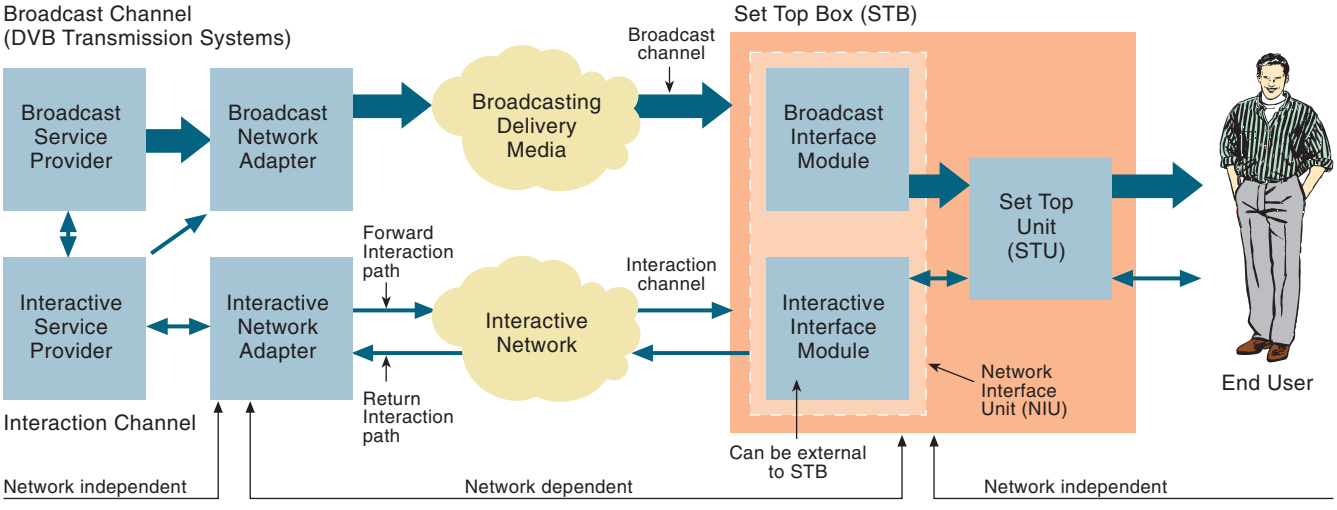


were identified in CM, the "Prosumer", the "Corporate" and the "Consumer". The main target will first be the "Prosumer", meaning "Professional consumer", which is a user with the need for broadband, high-quality services and the financial capacity to invest in relatively costly niche-market-equipment. The typical prosumers are home offices, small media or graphical design offices, medical or educational centres. The corporate will represent a larger group of users behind one terminal, typically with a LAN connected to the RCST (RCS Terminal). The usage of the channel will tend towards more symmetric behaviour, with the meshed network as the ultimate architecture. The consumer will probably be the last profile to feel the need and to be able to afford this kind of equipment, but a fast development in technology, together with increased needs for capacity and enhanced services makes the consumer a realistic user in the future.

In addition to defining the three user profiles, the CR also gives the reference model and many specific requirements for the system. One of the requirements is that the DVB-RCS specification [2] shall resemble the other specifications as far as possible, in particular the DVB-S (DVB-Satellite [3]). The reason for this is the wish to help a quick understanding of the specifications, to enable reuse of technology, and hence to reduce time-to-market. The specification shall be frequency-independent. It shall enable reliable network- and user security mechanisms, and incorporate an efficient transport layer. Interfaces with other infrastructures, such as PSTN, ISDN etc., shall be possible, and flexible terminals shall permit dynamic frequency allocation. The CR also describes the target bitrates, services, bit error rates, prices and availability for each type of terminal. The DVB-RCS specification only aims at defining the network independent layers; the network management and the services offered are left for the network operators and service providers to define.

Figure 2 gives the general DVB return channel reference model. In this model, the interactive network is depicted as independent from the forward channel. Very often, however, the forward interaction channel, or forward signalling channel, is integrated in the forward transport stream

Figure 2 DVB's general reference model for interactive networks



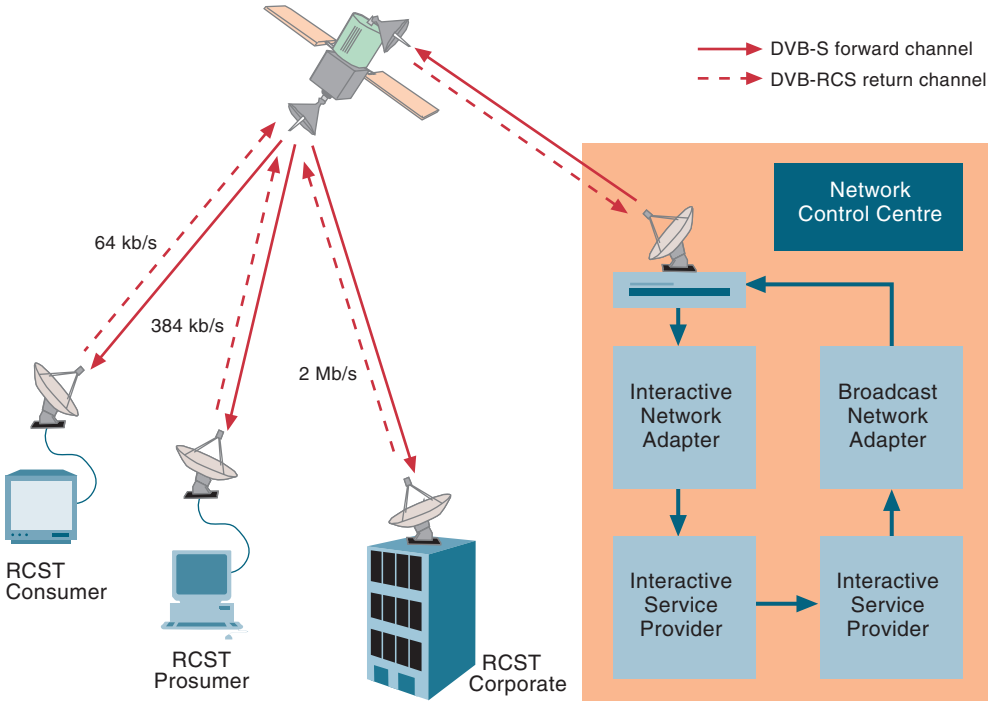


Figure 3 Simplified diagram of a network architecture for DVB-RCS systems

(TS). This is also the case in the DVB-RCS specification, where the forward signalling is part of the DVB-S TS. Figure 3 shows a simplified diagram of a network architecture. Actually, the DVB-RCS reference model is far more complex than this, however, the wish to indicate all possible network realisations may obscure the simplicity of the concept. Usually, several RCSTs will be connected to the interactive satellite network, consisting also of the satellite, an earth station and a network control centre (NCC). In Figure 3 the earth station antenna acts both as a feeder for the forward path and as a gateway for the return path. The NCC shall handle the synchronisation, give correction messages to the terminals and allocate resources.

One of the main challenges for the DVB-RCS group has been to specify inexpensive terminals. The CR indicates ex-factory prices in the order of 1,000 Euros, 3,000 Euros and 50,000 Euros for Consumer, Prosumer and Corporate terminals respectively. Cost limitations will also imply EIRP limitations and possible use of sub-optimal components such as nonlinear amplifiers. Of course, the cost is not the only limiting factor for the EIRP, regulatory rules have to be respected as well. In addition, the satellite channel is noisy, and use of Ka-band transmission for the return channel may give less known effects of multipath fading. As soon as interactive services are considered, the delay becomes a matter of concern with implications on several levels ranging from synchronisation, log-on algorithms to the delay perceived by the user after having made a request. This aspect highlights the need for efficient transport mechanisms, a need to be balanced against the contradicting need for flexibility. The NCC is in charge of the network control, which will include several RCSTs, but maybe also several satellites, feeders, gateways and even several networks. The RCST network to manage is a multipoint-to-point structure, far more complex to administrate than the opposite, the point-to-multipoint.

The NCC is thus in charge of the control of every RCST in the network as well as the network as a whole. A terminal will log on after having received general information by "listening" to the forward link. The information given there is on the status of the network, and most important, the forward link provides the Network Clock Reference (NCR). When the RCST has obtained synchronisation with the NCR, it will use one of the designated slots (indicated in the forward channel) for log-on request in a slotted-aloha manner. If the terminal is successful in this request, the NCC will forward various tables containing general network and terminal specific information. The specific information is about necessary frequency, timing and power level corrections to be performed by the terminal before transmission starts. These tables will also indicate the resources allocated to the terminal, and it is possible to ask for different services or increased capacity during transmission. The NCC has the possibility, with certain intervals, to correct the transmission parameters of the RCST, and if something goes wrong during transmission, the NCC shall also have the possibility to force the log-off of the RCST. The continuous signalling from the NCC is provided according to MPEG-2 SI [3].

The DVB-RCS specification is also restricted to the indoor unit, i.e. the signal processing between the source encoder and the IF conversion. The outdoor unit (the radio frequency part) will be specified by ETSI in [4]. The DVB-RCS physical layer contains specification of timeslots and frames organised in super-frames. The sequencing is controlled by means of the NCR, the access method is MF-TDMA (Multiple Frequency Time Division Multiple Access). Otherwise, the specification contains energy dispersion, two types of channel codes (concatenated Reed Solomon/convolutional coding and Turbo-codes), prefix emplacement, Nyquist filtering and QPSK modulation, most of which is well known from the DVB-S specification.

Conclusions

Many satellite operators have shown their interest in the return channel via satellite technology, and concrete plans exist for operation of such services in the near future. The prosumer market has been evaluated to have a potential market of some millions only in Europe, and as soon as higher volumes of terminals are produced, reasonable prices for the consumer market will be reached.

The DVB-RCS specification has been treated and approved in the TM, the CM and the Steering Board in DVB, and was sent ETSI in March 2000.

References

1 *DVB* (2000, June 19) [online] – URL: http://www.dvb.org/

2 *DVB Blue Book* A54 rev. 1.

3 ETSI. *Digital broadcasting systems for television, sound and data services; Framing structure, channel coding and modulation for 11/12 GHz satellite services*. Sophia Antipolis, 1994. (ETS 300 421.)

4 ETSI. *Satellite Earth Stations and Systems (SES); Harmonised EN for Satellite Interactive Terminals (SIT) and Satellite User Terminals (SUT) transmitting towards satellites in geostationary orbit in the 29.5 to 30.0 GHz frequency bands covering essential requirements under article 3.2 of the R&TTE Directive*. Sophia Antipolis, 1999. (ETSI EN 301 459 V.1.1.1 Draft.)

Kaleidoscope

Tore Olaus Engset's Wave Mechanical Discussion of the Hydrogen Atom

KRISTOFFER GJÖTTERUD AND BJØRN JENSEN

Kristoffer Gjötterud (69) is Associate Professor in Physics at the Department of Physics at the University of Oslo. He has published and lectured for a wider public on physics, epistemology and the relation of science and Jewish/Christian belief. He is a Human Rights activist.

Kristoffer.Gjotterud@fys.uio.no

Bjørn Jensen (35) is Associate Professor at the Sogn og Fjordane College, Faculty of Science. He obtained his Dr.Philos. degree in theoretical physics at the University of Oslo in 1997. He has been doing basic research on various subjects in theoretical physics.

bjorn.jensen@hisf.no
bjoje2@frisurfe.no

Tore Olaus Engset published six articles in *Annalen der Physik*[1)], all with the title "Die Bahnen und die Lichtstrahlung der Wasserstoffelektronen". The articles were received by the journal 28 June 1926 [a], 18 August 1926 [b], 19 August 1926 [c], 21 October 1926 [d], 15 February 1927 [e] and 26 October 1927 [f]. These works are the first theoretical studies, made by a Norwegian, of atoms on the basis of Erwin Schrödinger's paradigm setting work "Quantisierung als Eigenwertproblem"[2)] published in the same journal and received 26 January 1926 [1] and with an important note added in proof dated 28 February 1926. Before Engset published his first work, Schrödinger had published three more works[3)] received 23 February 1926 [2], 18 March 1926 [3] and 10 May 1926 [4]. The works [2] and [4] have the same title as the work [1]. The work [3] has the describing title: "Über das Verhältnis der Heisenberg-Born-Jordanschen Quantenmechanik zu der meinen". In the work [2] Schrödinger discusses the Hamiltonian analogy between mechanics and optics, and he applies his theory to the description of Planck's oscillator and to a fixed and a free rotator. In the last work mentioned here Schrödinger developed a perturbation theory which he then applied to the so-called Zeeman-effect. He also announced the necessity of a correct relativistic formulation of the Hydrogen atom within the framework of the new wave mechanics in order to handle the intrinsic spin and the magnetic moment of the electron. The magnetic moment of the electron had already been discussed by Uhlenbeck, Goudsmit, Langevin and Pauli.

Engset is leaning heavily on the mathematics in Schrödinger's work [1], but on some important points he does it his way. We shall here discuss those points.

Where Schrödinger derives his time independent, now well known, second order differential equation, hereafter referred to as the Schrödinger equation, by a variational principle applied to the energy equation

$$H\left(q, \frac{\partial S}{\partial q}\right) - E = 0,$$

with $S = K \ln \psi (K = \hbar)$ and where the function ψ is to be varied, Engset applies integration by parts to the same equation. Both have the problem to do away with a surface integral; Schrödinger with the integral

$$\int df \delta\Psi \frac{\partial \Psi}{\partial n} \tag{1}$$

and Engset with the integral

$$\int df \Psi \frac{\partial \Psi}{\partial n} \tag{2}$$

where $\frac{\partial \Psi}{\partial n}$ is the normal-derivative to the surface f. Notwithstanding the fact that the physical meaning of ψ is very obscure in both papers, Schrödinger treats ψ in a way that is mathematically remarkably close to the later mathematical understanding. Schrödinger brings about the quantisation of ψ via his variational principle above supplied with additional explicit mathematical requirements on ψ, which he takes to be "real over the whole configuration space, unique valued, finite, continuous and twice-differentiable". Schrödinger applies his theory immediately to the Kepler motion demanding ψ to satisfy the Schrödinger equation along with the stated boundary conditions. Schrödinger states explicitly that the integral (1) extends over an infinitely distant closed surface.

It is clear from the discussion in his paper [a] that Engset is trying to interpret ψ as a quantity describing the electron in its Kepler motion which he sees as a classical motion. In his own wording: "We assume that the dimensions of the electron are very small in comparison with its distance from the positive nucleus. Moreover, we consider the orbital motion, as well as the radiation from the electron, to be determined by the function ψ in the electron itself and in its immediate neighborhood". As for his integral (2) Engset comments "n is the normal of the fictitious spherical surface of the electron". In paper [c] he states that "... where f is the surface of the electron ... we get the important differential equation $\Delta\Psi + \frac{2m}{k^2}\left(E + \frac{e^2}{r}\right)\Psi = 0$

---

1) *[a]* Annalen der Physik, *80, (823), 1926; [b] ibid. 81, (572), 1926, [c] ibid. 82, (184), 1926; [d] ibid. 82, (143), 1927; [e] ibid. 82, (1017), 1927; [f] ibid. 84, (880), 1927.*

2) *[1]* Annalen der Physik, *79, (361), 1926.*

3) *[2]* Annalen der Physik, *79, (489), 1926; [3] ibid 79, (734), 1926, [4] ibid 80, (437), 1926.*

for the inside and on the surface of the electron". Engset mentions again the demand (2) = 0, but he does not discuss further which consequences this condition will have for ψ, and does not give any hint as to how this condition could be fulfilled. In paper [c] however, he does make some interesting remarks on the properties of ψ: "that the size of the electrons ... is also very small as compared with the possible existing extension of the period of spatial fluctuations of ψ. ... ψ is a function of the coordinates of the whole of space with the center of the nucleus as its origin". From these remarks it is obvious that Engset is struggling with the very meaning of ψ and that his mathematical conception of ψ is not consistent. This is further underlined by his remark (paper [c]): "The action *S*, which by definition can never be negative, ... has the immediate consequence that the possibility that

$$\psi < 1$$

can be excluded from the calculation". This remark is very strange in the light of Schrödinger's normalisation condition on ψ

$$\int \Psi^2 d\tau = 1 \qquad (3)$$

stated in the addendum to his paper [1]. Since Engset refers to Schrödinger's paper [1] it is very difficult not to assume that this must have been known to him. Nonetheless, he discusses the problem in his own independent way and he does not comment on Schrödinger's condition (3). The mathematical inconsistency between his conditions for ψ and the boundary conditions that Schrödinger imposes on ψ does not seem to bother him. This is a pity, since it is the boundary conditions, imposed by Schrödinger, which bring about the quantisation of the bound states for the Kepler motion.

Schrödinger discusses the Kepler motion in 3D, supposes that the dependence of ψ on the spherical co-ordinates is given by $\psi = \chi(r) \cdot \Phi(\theta, \phi)$, and gets for $\chi(r)$ the differential equation:

$$\frac{d^2\chi}{dr^2} + \frac{2}{r}\frac{d\chi}{dr} + \left(\frac{2mE}{K^2} + \frac{2me^2}{K^2 r} - \frac{n(n+1)}{r^2}\right)\chi = 0 \quad (4)$$

In a discussion of today, we will write the admissible solutions to (4) as:

$$\chi_{\ln}(r) \mu e^{-\aleph r} \cdot (2\aleph r)^n \cdot L_{l-n-1}^{2n+1}(2\aleph r) \qquad (4a)$$

$\aleph = \sqrt{\dfrac{-2mE_l}{K^2}}$, L_{l-n-1}^{2n+1} is an associated Laguerre polynomial of degree $(l - n - 1)$.

The energy eigen-value corresponding to the eigen-function (4a), as given by Schrödinger, is

$$E_l = -\frac{me^4}{2K^2 l^2}$$

which is the well known expression for the Bohr energy levels of the Hydrogen atom with *l* as the principal quantum number and where, as given by Schrödinger, *n* is an integer < *l*.

Engset tries to reduce the quantal Kepler motion to a classical motion in a plane. To achieve this he chooses cylindrical co-ordinates, and assumes that $\psi = \chi(r) \cdot e^{in\phi} \cdot f(z)$. Engset gets for $\chi(r)$ the differential equation

$$\frac{d^2\chi}{dr^2} + \frac{1}{r}\frac{d\chi}{dr} + \left(\frac{2mE}{K^2} + \frac{2me^2}{K^2 r} - \frac{n^2}{r^2}\right)\chi = 0 \quad (5)$$

under the assumptions that $z = 0$, and

$$\frac{d^2 f(z)}{d^2 z} = 0. \qquad (5a)$$

Since the physical system to be described consists of the electric interaction of small spherical charges the boundary conditions of the problem are not apt for cylindrical co-ordinates. This should suffice to conclude that the condition (5a) is not fulfilled, even at $z = 0$. Then we must conclude that (5) does not describe the quantal Kepler motion as described by Schrödinger.

Suppose that Engset's approach for $z = 0$ does give a correct answer. Then it should be possible to get the condition (5a) for the simplest of the eigen-functions (4a), which corresponds to the choice $l = 1$ and $n = 0$:

$$\chi_{l=1 n=0}(r) \propto e^{-\aleph r} \qquad (6)$$

From (6) it is clear that ψ cannot be factorised as supposed by Engset, and that we will always have $\dfrac{d^2\Psi}{d^2 z} \neq 0$, i.e. (5a) can never be fulfilled.

We shall here refrain from discussing which physical system would be described by Engset's approach. What we do know is that it is not the Hydrogen atom.

However, the factor $e^{in\phi}$ in Engset's wave function could be correct. On the assumption that $l = n \geq 1$, Engset relates [a] the quantum number *n*, though not explicitly stated, by applying Kepler's Second Law, to ν which he defines as "the number of periods made, when the image point and the electron make one complete cycle, respectively in the circular orbit and in the elliptical orbit, ...". Engset argues that "the number of *periods* made, when the image point and the electron make a complete cycle, respectively in the circular and in the elliptical orbit, is thus

equal to $^1/_2\, n$."[4] Engset derives the number N of cycles per unit time in the state given by the quantum number n with the help of the virial theorem for the classical Kepler motion and the energy eigen-value as given by Schrödinger:

$$N = \frac{4\pi^2 me^4}{n^3 h^3}$$

The crucial point, for which Engset does not give any further arguments, is to identify the quantity $^1/_2\, nN$ with the frequency of the radiation emitted "when the electron from a spatially distant point 'jumps' into the stable orbit that is determined by n." By doing this Engset arrives at the frequency condition, already established for a one electron atom by Niels Bohr[5]:

$$\nu = \frac{1}{2} nN = -\frac{E}{h} = \frac{2\pi^2 me^4}{h^3 n^2} \qquad (7)$$

Bohr, in his first famous paper published in 1913[6], makes the following principal assumptions:

"(1) That the dynamical equilibrium of the systems in the stationary states can be discussed by the help of the ordinary mechanics, while the passing of the systems between different stationary states cannot be treated on that basis.

(2) That the latter process is followed by the emission of a *homogeneous* radiation, for which the relation between the frequency and the amount of energy emitted is the one given by Planck's theory."

Bohr makes it very clear that "the second assumption is in obvious contrast to the ordinary ideas of electrodynamics". Engset does not refer to Bohr or to any other literature discussing in depth the fact that the frequency of the radiation emitted is half the frequency of revolution of the electron in its final state when trapped from a great distance. The factor $^1/_2$ between the frequency of the radiation and the revolution of the electron must however have bothered Engset since he returns to this factor in the beginning of paper [b], but without adding anything to the understanding of it.

In his next paper [b] Engset addresses in more detail the phenomena connected with the transition "from one stable orbit to another stable orbit" in the Kepler motion. He makes the interesting remark: "In order to make a first step in this for us obscure area of research, we have assumed ...". In his further discussion Engset does not relate to Schrödinger's 4th Communication[7], received by *Annalen der Physik* 21 June 1926, where Schrödinger both introduces his time dependent differential equation as well as discusses the radiation process using perturbation theory. This paper might not have been available to Engset at the time of writing his second paper. With a reference to his paper [a] Engset states: "Nothing is here said about the time dependence of the radiation – whether it is damped, or undamped and then suddenly broken off. Similarly, under this restriction of the problem, nothing could be said about phenomena connected with the transition from one stable orbit to another stable orbit." Engset then enters a discussion of the charge distribution of the electron itself on a classical basis. Struggling with the cause of the radiation following a transition from one stable orbit to another, in both of which the electron does not radiate, Engset proposes a model where a change of the charge distribution of the electron itself, during the transition, does generate the radiation. A remark in the paper (b) suggests that Engset did not accept that an electron in a stable orbit does not radiate. We cite his remark in its original phrasing: "– dass die Strahlung während der Bewegung des Elektrons in der stabilen Bahn ausgesandt und durch Schwingungen der elektrischen Ladung desselben verursacht wird".[8]

Engset describes the charge oscillation on the electron by a classical second order differential equation for the time dependence of "the excess charge on the first half surface relative to the equilibrium charge":

$$\frac{d^2 q}{dt^2} + 2\beta \frac{dq}{dt} + \left(\beta^2 + k^2 \varepsilon_{mn}^2\right) q = 0 \qquad (8)$$

The constant β is the resistance to the current in the electron, $k = {}^1/_h$ and the energy difference ε_{mn} between the m and the n orbit is introduced

---

[4] *The mapping of the motion of the electron in its Keplerian orbit on a circle in the orbital plane, with the semi-major axis as radius and with the centre in the positive nucleus, is discussed by Jacob Nielsen in his* Lærebog i Rationel Mekanik II Dynamik*, Jul. Gjellerups Forlag – København 1945 based on lecture notes 1934.*

[5] *On the Constitution of Atoms and Molecules,* Phil. Mag. S*, 6, (151.1), 10, July 1913.*

[6] *ibid. p 7.*

[7] Annalen der Physik*, 81, (109), 1926.*

[8] *"... that the radiation emitted during the motion of the electron in the stable orbit is caused by the oscillation of its electric charge."*

ad hoc. In Engset's model (8) there is a damped oscillation of the surface charge of the electron, with a frequency $\nu_{nm} = \frac{\varepsilon_{mn}}{h}$, when the electron makes a transition from the orbit *m* to the orbit *n*. This was, already in 1926, very far off the mainstream understanding of the quantal radiation process in the atom. Therefore we shall not make further comments on Engset's following three papers [d], [e] and [f].

Summary and Evaluation

In his papers published in *Annalen der Physik* during 1926 and 1927 Engset makes references to Schrödinger [1] in paper [a], to M. Abraham, *Theorie der Elektrisität* (1905) in paper [c] and to E. Waetzmann, Zur Theorie der Kombinationstöne, *Ann. d. Phys.* 24, (68), 1907 in paper [e]. There are no other references.

Though it is easy to see that Engset has been strongly influenced by the first wave mechanical work published by Schrödinger, he chose his independent approach from the very outset. He quotes the correct energy eigen-values for the Hydrogen atom, but he cannot, by his approach, have derived them. In spite of Schrödinger's explicit demand on ψ to be quadratic normalised to unity when integrated over whole space, he maintains that ψ should never be less than one. Schrödinger also stated that the quantum number *n* "which gives the order of the spherical harmonics appearing in the solution, may then always be given only values smaller than *l*", where *l* is the principal quantum number. Engset nevertheless discusses the case $l = n \geq 1$.

It seems to have been very important to Engset to understand the dynamics of the Hydrogen atom by a reduction of the wave mechanical description of Schrödinger to classical concepts with which he is more familiar. This is evident when he approaches the physics of radiation from the atom by his model of a charge oscillation on the electron itself. The quantal properties are again not derived, but introduced ad hoc.

The critical remarks we have had to make on Engset's works must be fully shared with the editors of *Annalen der Physik*, since they obviously did not give Engset any feedback and corrections prior to publication of the papers.

What is remarkable however is that Engset, so quickly after Schrödinger's introduction of his wave mechanics, got interested in the field and, as the first Norwegian, tried to do something about it in his own independent and original way. We have learned[9)] that Egil A. Hylleraas, in a lecture at the Department of Physics at the University of Oslo, probably in March 1963, mentioned that except for the interest for the new quantum mechanics shown by the Director of The Norwegian Telecommunication Authority, Tore Engset, the knowledge of this new field was very poor in Norway at that time. This further adds to the picture of a man struggling with difficult problems at the forefront of the science of his time.

Engset was also fully aware that Schrödinger had entered a scientific *terra incognita*. He was also aware that his own approach was preliminary and incomplete. He ends his paper [c] with the phrase: "Hoffentlich wird sich bei eingehender Betrachtung der möglichen Vorgänge im Electron und im umgebende Felde Zeigen, dass die forstehende Theorie, durch die *Erfahrung* geprüft und matematisch vollständiger (durch Vectoranalysis und Relativitätstheorie) behandelt, zur *Klärung* der sich aufdringenden Fragen beitragen wird. Denn Abraham hat gesagt: "Für eine Theorie ist die Eroberung einer neuen Provinz stets ein Unternehmen, das nur der Erfolg rechtfertigen kann"."[10)]

The development of physics has proven that Engset was on the wrong track. It should not be forgotten, however, that it does show courage when one takes the risk of being wrong. Tore Engset should be honoured for being the first Norwegian to recognise and to be inspired by the paradigm giving work of Erwin Schrödinger on wave mechanics.

---

9) *Private communication, professor Leif Veseth, Dept. of Physics, University of Oslo.*

10) *"Hopefully it will be demonstrated, by a careful consideration of the possible processes in the electron and the surrounding field, that the present theory will contribute to a clarification of urgent questions through the tests given by experience and by a more complete mathematical treatment (by vector analysis and the Theory of Relativity). For Abraham said: «For a theory the conquering of a new province is always an enterprise that can only be justified by its results.»"*