Charging and Pricing in Multi-Service Wireless Networks

Henrik Franzén

Stockholm
2001

Master Thesis
Department of Microelectronics and Information Technology
Royal Institute of Technology

Abstract

This master thesis analyses different factors that affect the process of charging and pricing for services and applications in UMTS networks. The report contains a study of future market players within the area of mobile communication and a discussion of enabling network technologies. It investigates what services seem to be the most interesting in a UMTS context, the way customer demand is predicted as well as mobile traffic and QoS related to the provisioning of the services/applications. Moreover, there is a description of a possible future charging framework and a brief review of the UMTS business case.

Today, network owners are the most profitable actors within the business of mobile communication. However, the role of content providers is believed to become more prominent for UMTS. This is supported by empirical studies of the Japanese market and the fact that content generates a more attractive margin than ordinary voice, which in turn is a vital argument when it comes to regaining costs for the infrastructure investments.

The usage of services and applications must be charged for in order to balance demand and supply of the scarce UMTS radio spectrum. The increased requirements for QoS also support this assumption. The most suitable basis (e.g. time, volume) for charging will probably vary from one case to another. However, the evolution towards "all IP networks" partly includes an increased focus on the volume of transmitted data.

The most crucial point is perhaps not how to charge for the network activity, but how to price it. The price per bit in UMTS is believed to be about the same as in GPRS. This certainly does not favor resource-demanding services such as streamed applications. Simultaneously, there are services that consume insignificant amounts of resources (e.g. emails), which are almost for free but which contribute notable consumer value. Thus UMTS prices will very much be a question of balancing perceived customer value against the costs of provisioning. Here, price discrimination is useful in order reveal people's propensity to spend money on communication services. Of course, the competitive rules between different market players also affect the final price.

Predicting the market demand curve tend to be difficult. There are so many parameters that ought to be taken into consideration. Issues like age, gender, IT literacy, profession, price sensitivity and the personal budget, all matters to the individual demand. Besides, human actions are often unpredictable and irrational, which makes the task even more complicated.

In summary, determining the right charging and pricing schemes for mobile services and applications is no easy assignment. The final result will be a solution that is dependent on several determinants, which are hard to identify in advance.

Acknowledgements

This master's thesis is my final work for the Master of Science in Industrial Engineering and Management at the Royal Institute of Technology (KTH) in Stockholm. It was performed at Ericsson Research (Switchlab), under the supervision of Carl-Gunnar Perntz and Tord Westholm. I would like to thank both of them for their help and assistance during the entire work. I also would like to express my appreciation to Dr. Terje Jensen and Dr. Ragnar Andreassen at Telenor FoU, for valuable information and fruitful discussions. Finally, I would like to direct my acknowledgement to Professor Gunnar Karlsson for his support and for being my examinator at KTH.

Table of Content

1. Introduction

1.1 A Brief Introduction to the Area of Interest

The rollout of UMTS (Universal Mobile Telecommunication System) services implies new wireless communication facilities. The mobile access networks will enable transmission speeds of up to 2 Mbps, enabling services and applications with real-time characteristics. The charging and pricing of these activities is crucial for the entire UMTS business case.

Notwithstanding, this turns out to be a complex and complicated task, since there are so many simultaneous claims, both technical and economical.

Charging will influence the network utilization and is closely linked to the networks' ability to guarantee QoS (Quality of Service). This in turn concerns questions related to network dimensioning. Moreover, different tariff structures will have different effects, depending on the underlying network characteristics.

The intention is to maximize network utilization and profits at the same time. Incomes must cover for the costs associated with the UMTS infrastructure and at the same time correspond to customers' willingness to spend money on mobile communication. Charging must not be studied in isolation, but in a context including enabled services / applications and customer demands.

Hence, the problem is not only a question of how charging and pricing are actually tried out. In order to be able to approach this area, enabling factors turns out to be equally important. Some of them are further discussed in the thesis.

1.2 The Aim with the Report

Questions related to the pricing and charging of third generation applications and services are highly relevant for the entire business case. In this respect, Ericsson has a central role as a promoter of the equipment they are developing and which is needed in order to realize the functions of the networks. This report aims at supporting Ericsson with knowledge related to the operators' businesses, which might be valuable when it comes to the actual demand for network equipment.

1.3 Research Problem

The research problem is about describing and identifying different factors and parameters, which must be considered when determining how to price and charge mobile services in UMTS. The objective of this work is to present and discuss basic research within the area of mobile communication and to investigate what questions might arise when deciding on reasonable charging and pricing structures.

1.4 Limitations

- There are numerous factors, which affect the way services and applications are priced and charged. This work does not cover them fully.

- The study is mainly focused on the European market, but information from other parts of the world is considered in the presentation. This constitutes a significant weakness, since cultural differences often make it hard to draw precise conclusions concerning local or regional conditions. What is true in Japan does not necessarily have to be true in Sweden or in Norway, for instance.

- I will restrict my presentation to one single charging architecture. It was established within the M3I (Market Managed Multi-Service Internet) project and serves as a useful and conceptual model of charging in real time, including multiple providers.

- The supply of reliable information about strategic pricing and charging policies is poor. Consequently, parts of the presentation are made up of qualitative assumptions and predictions.

- A high level of abstraction characterizes the section about customer demand. This is due to the shortage of examples from real life.

- Most experiences with network charges so far come from fixed networks. Often, there are dissimilarities that make comparisons between fixed and mobile networks hard to perform. Thus, the information must be seen from a "mobile perspective".

- The study is mainly written much from a UMTS operator's point of view.

1.5 Structure of the Report

The report will cover a wide theoretical area, starting with a background description of topics, which in some way concerns the area of charging and pricing. The purpose is to introduce the reader to the subject in question.

Chapter three investigates the market for mobile communication. The objective is to give a picture of what actors there are and their interdependencies. The next chapter deals with the evolution of the network infrastructure, predominantly "all IP" (Internet Protocol) networks.

Chapter five investigates what services seem to be most interesting in the eyes of the consumers. This is followed by a chapter, which looks at the concept of quality of service. In chapter seven, there is a discussion about components crucial for the market demand curve. Thereafter, I will look into a simple model used to describe mobile traffic on a high level of abstraction.

The last three chapters are dedicated to pricing, charging and an overview of the UMTS business case.

1.6 Methodology

I have spent my time studying written information within the area of charging and pricing and other topics closely related to the subject. Much of the information is found in scientific research papers, master theses and different kinds of technical magazines and newspapers. I have had regular meetings with people at Telenor and also gained valuable knowledge about the subject during several informal discussions with employees at Ericsson.

1.7 Switchlab

Switchlab, which is part of Ericsson Research, was established 10 years ago and is physically, located in Kista outside Stockholm. The main activity concerns research about real-time routing, ad hoc networks and QoS. Switchlab is also involved in standardization issues, IP load control, header compression etc.

1.8 Definitions

This section presents some conceptions, frequently used later in the report. They coincide much with the definitions listed in (M3I 2000b).

Accounting:
Summarized information (accounting records) in relation to a customer's service utilization. It is expressed in metered resource consumption, e.g. for the end-system, application, middleware, calls or any other type of connection.

Billing:
Collecting charging records, summarizing their charging content, and delivering a bill or invoice including an optional list of detailed charges to a user.

Charges:
Charges determine the amount of monetary value that needs to be paid for particular resource utilization. It is contained in a charge record.

Charging:
The overall term "charging" utilized as a summary word for the overall process of metering resources, accounting their details, setting appropriate prices, calculating charges, and providing a fine-grained set of details required for billing. Billing itself is not included in this definition.

Pricing:
The specification and setting of prices for network resources.

Service:
A service enfolds autonomous and network dependent tasks needed for application execution. An application typically employs several and presumably distributed services to provide full functionality.

Tariff:
The algorithm used to determine a charge for a service usage.

Metering:
Determining the particular usage of resources within end-systems or intermediate systems on a technical level, including quality of Service (QoS), management, and networking parameters.

2. Background Studies

The Evolution Towards UMTS

GSM

Mobile data services of today make use of the GSM (Global System for Mobile Communication) platform, which enables SMS (Short Message Services) and circuit switched data. The SMS normally consists of a text message of maximally 160 characters routed via the control channel. Circuit switched data is transferred on the normal traffic channel at the rates of 13 Kbps (encoded voice) or 9.6 Kbps (data), (Prasad 1999).

The user pays for the duration of the session, and the connection set-up is the same irrespective of what type of information that is in transfer. However, since the penetration of these services increases, the allocation of scarce radio frequencies is getting ineffective.

The majority of the established network operators have invested tremendous amounts of money in GSM supporting infrastructure. Hence, they demand a smooth and cost-effective evolution towards 3G (Third Generation) networks in order to reuse the existing equipment (Rouz et al. 1999).

HSCSD

HSCSD (High Speed Circuit Switched Data) seems appealing to real-time services, such as videoconferences. A flexible air interface resource allocation makes it possible to combine the mobile operators' strategies and the users' needs. HSCSD enables the user to transmit and receive data on more than one time slot at the same time, which multiplies the transmission speed by the number of timeslots used. Moreover, new effective coding schemes allow data rates of up to 14.4 Kbps per channel (Andersson & Wirde 2000).

Resources could be allocated either dynamically (non-transparent) or statically (transparent). By using a technique called Dynamic Downgrading Priority, timeslots are taken from non-transparent calls, which allocate more than one timeslot and given to incoming non-HSCSD calls. As a result, subscribers could not be guaranteed a uniform and stable throughput.

For a transparent HSCSD call, the number of allocated timeslots is fixed. Unfortunately, this is associated with set-up and hand-over problems. The number of requested time slots in a new network cell might be allocated already, which results in the dropping of calls (Andersson & Wirde 2000).

HSCSD services may be charged for in different ways, including:

- The number of timeslots used.
- Flat rate, irrespective of usage.
- The number of bits transferred during the call.

The migration from the GSM to HSCSD seems to be easy and requires only minor software upgrades (Andersson & Wirde 2000). However, the user will have to invest in a new mobile terminal.

GPRS

GPRS (General Packet Radio Service) means overlaying a packet based air interface on the existing circuit switched GSM network. Packet based traffic implies that information is segmented into packets, which take different routes through the network. At the receiver, packets arrive arbitrarily, but are sorted in the same order as when they were sent.

This network technology admits an increased number of users simultaneously and improved bit rates. It also means efficient usage of the radio spectrum, since traffic from different subscribers is statistically multiplexed. Thereby, the need to provide capacity that is only used at peak hours will be reduced.

Parts of the messages, previously sent using SMS, will probably migrate to GPRS, but the need for SMS as a complementary bearer service will remain. It is unlikely that network operators will allow too many time slots to be consumed by a single GPRS subscriber (http://gsmwold.com).

Applications requiring relatively high-speed transmissions, like high quality video, are not enabled until the introduction of EDGE (Enhanced Data for Mobile Evolution) or UMTS. The fact that GPRS packets are exposed to transit delays enforces the need for EDGE.

An early development of GPRS would be an excellent opportunity for operators and service providers to investigate the feasibility of new applications before the introduction of UMTS. Services like e-mail will be more flexible to use, since there is no longer any need for the connection set-up. Transmission speeds of up to 171.2 Kbps are planned, but the maximum data transfer will eventually depend on the infrastructure (Prasad 1999).

Volume based charging schemes are applicable so that the user can stay online all day and pay for the actual volume of transmitted data.

EDGE

The allocation of multiple time slots has been the primary way to increase data rates in GSM networks. EDGE can be viewed as an extension of the GSM standard, just like HSCSD and GPRS (Prasad 1999).

EDGE uses a new modulation technique in order to reach higher data rates on the radio interface. It can be used for circuit switched (with HSCSD) or packet based (with GPRS) technologies. The enabled bit rate is 48 Kbps per channel, which corresponds to 192 Kbps in each direction, in case four channels are used (Andersson & Wirde 2000).

The investments required for implementing EDGE include new EDGE transceiver units and software upgrades. Furthermore, terminals must be developed, which support the new modulation technique (Andersson & Wirde 2000).

UMTS

UMTS will allow delivery of voice, graphics, video and broadband information of all types. The access/radio technology for these services will be WCDMA (Wideband Code Division Multiple Access), supporting a combination of packet and circuit switched data. This will allow a customer to join a videoconference, download data files and access web pages, simultaneously. Some of the technical benefits achieved thanks to WCDMA compared to the GSM system are presented in Table 1 (UMTS Forum 1999).

Service flexibility	Effective usage of radio spectrum	Capacity & coverage	Economies of network scale
Each 5 MHz carrier enables a variety of services ranging from bit rates of 8 Kbps to 2 Mbps. Circuit and packet switched services are combined on the same channels.	Radio frequencies are efficiently utilized. The number of approved phone calls within a cell sector is increased.	The transceivers can handle eight times more voice compared to narrow band transceivers.	WCDMA can reuse the already existing GSM core network, which gives the existing operators the chance to build on the existing investments.

Table 1. Technical benefits associated with the WCDMA platform

Charging schemes in UMTS will open up for a new way of thinking. For instance, reverse charging and the charging of a party not participating in the call is under consideration. These circumstances demand for some kind of cost control implementation, where the user can limit the allowable usage. It should also be possible to charge all parties or only the originating party for so-called multi-leg calls (e.g. conference, forwarded or roamed calls) (ETSI 2000).

Trends on the Mobile Market

Actors

The number of actors on the market offering value added services will probably increase, and the multimedia services provider (MSP) will presumably constitute a key function in tomorrow's networks. It will purchase multimedia information from actors like TV program providers and software developers.

The MSP will make the information available in forms like Video-on-demand or interactive games, which is enabled for the customer via a fixed or mobile link on a multimedia server. High initial costs due to large investments, like an increased number of required base stations, could however lead to comparably high tariffs (UMTS Forum 1999).

If content is included in the distributed mobile multimedia it is likely that the provider expects payment for its value. Payments may be collected from the customers directly, but typically the operator carries this out. The actual flow of money between different stakeholders is still an open question. One can assume that a user demands both basic telecommunication services such as voice and a variety of value added services. These will be supplied both by network operators and value added service providers. The VASP (Value Added Service Provider) pays the operator for the resources consumed in order to satisfy the user. The customer on the other hand pays both the operator and the VASP. Hence a high level overview of payment flows could look like the ones depicted in Figure 1.

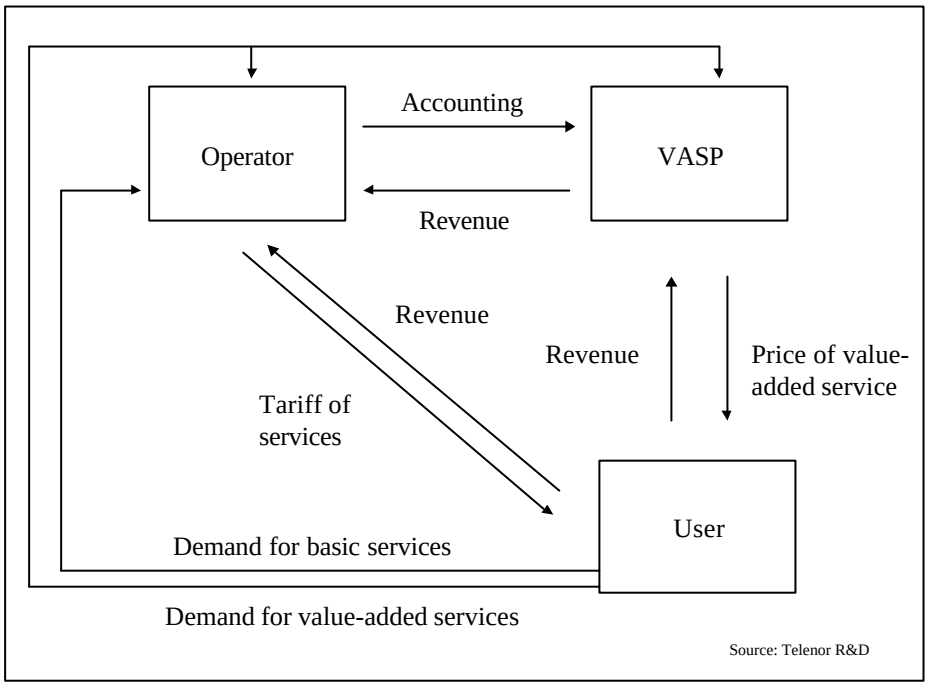

Figure 1. An illustration over money flows between different actors within the area of tele- and data-communication

Content may even be provided free of charge. In this case, including advertising generates revenue. Since content providers are curious about reaching the mass market, it is in their interests to minimize the costs of their services to the end users (UMTS Forum 1999).

However, only a minor piece of the total payments is shared with content providers. This is obvious within the cable TV business, where only one third of the total revenues goes to content, while the rest is transferred to the network operators (Odlyzko 2000). This implies that the network owner would get a completely new source of revenue if convergence moves delivery of content to the Internet.

Services

The wish to turn spare time into effective time both in a person's business and personal life, in addition to increased individual mobility, drives the development towards multimedia ser-

vices while on the move. The 60 % annual increase in the number of cellular phones each year in Europe supports this statement (UMTS Forum 1999).
Here are some additional trends, which may influence the creation of future services.

Social trends	Technology trends
Personal productivity.	Cost/performance trend in terminal components.
Need for personal security, due to an increase in crime.	Improved human-machine interface.
Demand for remote access due to flexible working practices.	Database and data compression technologies.
Demand for home entertainment services.	Development of effective usage of radio spectrum.

Table 2. General trends, which influence future mobile services

Future applications will include collaborating work systems, which enables "virtual project teams", stock checking, co-ordinated scheduling, fast access to digital libraries, virtual reality walkthrough of architectural designs etc.

In the future, the distinction between voice, video and data will become diffuse. We will probably witness a convergence between different communication sources, including entertainment, commerce and computing. So far, only fixed networks have been considered. Mobile multimedia will be a sub-set of multimedia services via fixed networks, which makes it important to define the fixed and the mobile multimedia markets in parallel. Fixed multimedia services today can be seen as good hints about future mobile multimedia attractions.

Mobile multimedia is divided into three generic types: medium, high and high interactive multimedia. Typical medium multimedia services are Intranet/Internet access, application sharing, interactive games and sophisticated broadcast and public information messaging. High multimedia services are for instance video and audio clips on demand, fast access and online shopping. Finally, video telephony, videoconferencing and telepresence serve as good examples of high interactive multimedia (UMTS Forum 1999).

Table 3 gives some examples of characteristics regarding different multimedia services according to UMTS Forum. The information in the table only serves the purpose of better understanding. The question marks signify that the characterizations are vague. This type of information is still valuable when dimensioning networks and when establishing suitable charging schemes. For example, asymmetric traffic behavior favors schemes based on the volume of transmitted data. Symmetric traffic is perhaps better charged for the time connected.

	Medium multi-media	High multimedia	High interactive multimedia
Asymmetric / Symmetric	Asymmetric	Asymmetric	Symmetric
Type of tariff (proposals)	Charged per MB?	Charged per MB?	Charged per minute?
Typical file size	0.5 MB	10 MB	
Delay sensitivity	Delay tolerant	Delay tolerant	Delay intolerant

Table 3. Assumed mobile multimedia characteristics

There are several outlines of scenarios regarding the future mobile multimedia market. It should be remembered though, that people tend to overestimate what can be done in one year and to underestimate what can be done in five or ten years. The mobile telephony industry is associated with a rapidly growing sector. However, it has taken about 15 years to reach the present level of usage (Odlyzko 2000).

The Traffic

The vision of UMTS is a convergence of fixed and mobile networks. Both circuit-switched traffic with guaranteed QoS and packet-switched traffic (both connection-oriented with guaranteed QoS and best effort) must be supported. Moreover, precaution must be taken to the asymmetry of the traffic that is caused by coming multimedia features.

Altmann et al. (1999b) analyses different types of traffic with respect to bandwidth over fixed networks. These are bulk traffic (FTP, streaming data), burst traffic (WWW) and interactive traffic (Telnet, X Windows). Bulk traffic is characterized by a small number of packets compared to the number of bytes transferred. The average packet size is larger than 1000 bytes. Packets of an average size smaller than 45 bytes represent interactive applications and the remaining traffic is classified as burst traffic.

Experiments show that burst traffic dominates (between 66-85 % of all traffic). However, as the supply of bandwidth increases, the proportions of both bulk and interactive traffic will increase.

Altmann et al. (1999b) also found that 62,5 % of all users in the experiments took advantage of bandwidths ranging from 8 to 128 Kbps at least at some occasion. The trials reveal little intra-user[1] variations in weekly mean expenditure, but that inter-user distributions of individual budgets and general usage are very heterogeneous. Price differentiation based on high or low volume users will however not be enough. Customers must be given the opportunity to switch between qualities more or less instantaneously.

In 2005 a significant proportion of all mobile traffic is assumed to be transmitted over a packet or cell-based network. Data could account for over 70 % of the total mobile traffic. Therefore, spectrum requirements and network design must be seriously considered. Tech-

[1] Intra-user expenditure concerns one single customer, while inter-user expenditure refers to the spending among different users (user groups).

niques like data compression and configurable radio interfaces will be important when demands on the spectrum increase (UMTS Forum 1999).

Tariff Structures

Overall Economics

Charging and pricing aim at maximizing the return on the network investments. Therefore, what is needed is a charging scheme that can reflect the true cost of service provisioning. Shaoyan & Chuanyou (1998) and Fishburn & Odlyzko (1999) discuss various costs, which could be derived to the operation of 3G networks. Here are some of them:

- Ongoing operational costs such as financial funds
- Salaries
- Welfare funds
- Repairing costs
- Cost for expandable pasts with low prices and service provision costs
- Depreciation of the network investments
- Network management and other overhead costs

Some of them will dominate over others. For example, the cost for depreciation will be much higher than in 2G networks, due to the huge infrastructure investments[2]. On the other hand, the costs for salaries will probably not be much higher than today.

Costs associated with repairing costs and network management will probably depend on factors such as QoS guarantees and interconnection agreements with peer networks etc.

Actually costs do play a dominant role for the operators' profitability. It seems as if the difference in revenues is caused more by different cost patterns than by different demand functions (Fishburn & Odlyzko 1999). The operators must be able to handle the transition from yesterday's resource scarcity and today's relatively high cost of local access to tomorrow's increased supply of high quality links.

Licenses

There are several ways UMTS licenses could be distributed and the UMTS Forum has listed several recommendations to national regulators. If the distribution includes money transactions, the sum of it should serve the purpose of being cost-recovery and nothing else. The high start-up costs will surely have considerable impacts on the UMTS business case, not the least in countries where license expenditures are tremendous. If license fees exceed administration costs, they could in fact result in direct negative effects on the deployment of UMTS in a large scale (UMTS Forum 1998). High up-front fees put serious strains on operators' balance sheets and require huge incomes. As a result, high prices might result in customers being locked out from the market.

[2] The network investment is not a cost. The networks are in operation and the cost is spread over several years. The annual depreciation however is a cost.

Below, some ways of assigning licenses are described, combined with typical characteristics associated with each of them (UMTS Forum 1998).

First come, first served	Comparative bidding (beauty-contest)	Auction	Lottery
The most widespread and long-standing method. It is appropriate if there is no scarcity of frequencies.	Pre-defined selection criteria are determined. This method could become complex and time-consuming.	Auctions might result in high up-front fees, which increase tariffs for the end customer. Auctions could also harm the competition because of incomplete information.	Lottery does not guarantee that awarded operators are competent enough. The method is seldom used in Europe, but it is quick and non-discriminatory.

Table 4. Different ways to regulate the radio frequencies

The goal of issuing licenses for spectrum usage is to ensure the appropriate service quality and to ensure the delivery of wireless broadband services. This seems to be a necessity, since new wireless services are expected to cause gaps between the demand and supply of radio spectrum. Here charging and pricing serve the purpose of controlling the usage.

Existing Charging Models

The existing billing models are adapted to POTS (Plain Old Telephone System) and the Internet. POTS charging is based on access (flat rate), location[3] and usage. Since the cost drivers for usage and access differ, the operators often use two-part tariffs[4]. Usage costs are traffic dependent while access costs vary with the number of subscribers and hence the charges consist of three major parts (Shaoyan & Chuanyou 1998):

- Installation and test charges
- A charge for connection with the network
- Lease charges

However, a marginal cost approach to pricing will not be functional for a telephone company since the break-even constrained optimum will not be satisfied without external subsidies or price discrimination. In practice prices have to be higher than optimal since the relationship between price, demand and capacity is constrained according to Figure 2 (Edén & Arvidsson 2000).

[3] Charging based on location is getting obsolete.
[4] That means, charges are based on access and usage for example.

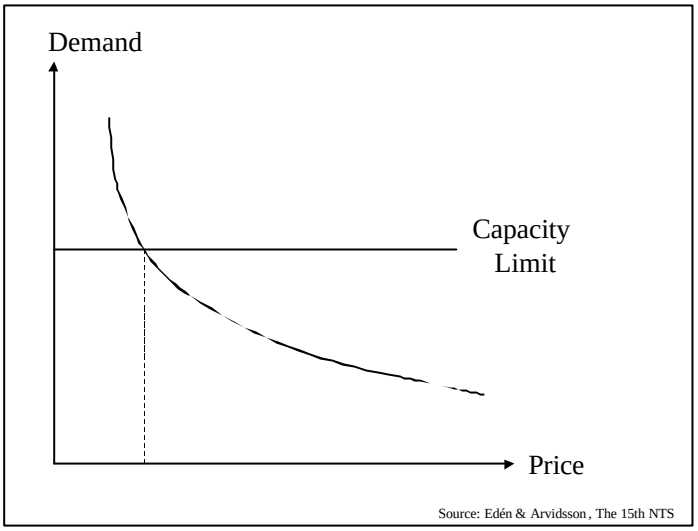

Figure 2. The price under a capacity constraint

The telephone market (in Sweden) has long been characterized by monopoly and charging has predominantly been based on rate averaging and cross-subsidization. Despite the heterogeneity in usage patterns between different customers, almost all user groups face the same prices.

Historically, long distance and business services have been priced disproportionately high in order to subsidize local and residential services. The cost for building and maintaining the local loop are cheaper for business customers, often located in densely populated areas. Thus it could be argued that telephone rates for business users should in fact be lower. Even low-cost users have been forced to subsidize high-cost subscribers (Sung & Cho 2000).

POTS charging schemes have the following characteristics. The models for mobile services are based on the same idea, but they are more complex.

- Costs are reflected
- Accounting is simple (a billing record is generated at the local exchange)
- A single QoS is supported

Internet charging is often a flat rate combined with some fee for connection time. Future schemes on the other hand require costs to be reflected in detail.

Today's schemes are not good for encouraging innovative usage of services. Unfortunately, time and distance based charging can be very inflexible (Rouz et al. 1999). For example, a WAP user is charged for the time spent online, not for the actual service[5]. The telecommunication market has rapidly turned into a competitive market. This impedes regulators and legislators from relying too heavily on uniform rates. Therefore, there have been attempts to

[5] Circuit switched networks.

price services more accurately according to their costs. Rate balancing has been successful in many developed countries considering the elimination of cross-subsidy from long-distance to local services. Besides, high-volume business users are sometimes offered price discounts.

Tomorrow, revenues per subscriber for some business segments will probably increase due to the substitution of fixed voice traffic against the use of value added services. However, many business users will probably generate comparatively low revenues due to the usage of wireless VPNs (Virtual Private Networks) at lower tariffs compared to public switched services (UMTS Forum 1999).

Usage Based Charges

3rd generation mobile network charges will almost surely depend on usage. This is motivated by the overall trend towards "all IP" networks, where the actual number of packets transferred is what really matters. Other reasons are the scarcity of radio resources and its classification into different traffic classes, which in turn delivers a multilevel QoS. In case flat rate charging is used, it would negate the benefits of this provisioning (Barnett 1998). According to Altmann et al. (1999a), flat rate charging implies over consumption of resources and loss of incomes from customers ready to pay higher fees.

Parameters associated with the session charging are connection-time, transmitted volume and file information indicators. These parameters could be expressed as in the following charging function (Botvich et al. 1997):

Fixed charge (*file content*) + E (*charge per unit time*) * Time + F (*charge per packet*) * Volume.

The scheme was evaluated both from a wholesale[6] and a retail[7] perspective regarding audit, practicality, usage sensitivity and predictability. This is illustrated in Table 5.

[6] See section 3.2
[7] See section 3.2

	Audit	Practicality	Usage sensitivity	Predictability
Wholesale charging	The counting of cells/packets is probably as reliable as information sent from a PSTN timer today.	More work is needed in order to make charging scalable. The number of different types of networks and actors will increase and then charging becomes a more complex issue.	Burstiness is not taken into account.	Financial responsibilities and the extent ghost traffic[8] exists needs to be clarified.
Retail charging	Volume based charging is similar to the wholesale approach. A full implementation on the Internet of an audit trial for time-based charging (only) seems hard to realize.	This is strongly dependent on the development of suitable Internet tools.	Improved flexibility can be attained thanks to the combined charging approach.	If there is a method of forecasting usage expenses at the application level, a high degree of predictability will be allowed.

Table 5. Charging scheme evaluation.

MacKie-Manson & Varian (1995) have proposed a two-part tariff consisting of fees for subscription and usage, which results in an efficient level of consumption as well an efficient level of capacity allocation. It also says that new firms will enter the market until the profits are driven to zero, in case there are no specific restrictions.

What will happen to the utilization rates during usage-based charges is unclear. With no usage price the resources are used more. What happens to the utilization will depend on the possible increase in capacity. The utility of a network service could be higher or lower than with usage based pricing, since there is more usage without prices, but also more congestion (MacKie-Manson & Varian 1995).

[8] Ghost traffic does not have anything to do with the service or application. Typical examples are connection set-up signaling and pinging.

Some charging schemes take congestion into account, but in order to reach the benefits associated with congestion charging, traffic must be of best effort type only (Barnett 1998). How to realize this method in a UMTS environment is still an open question. Perhaps, it could be useful in the unspecified service classes.

Congestion fees could be useful, since the operator becomes aware of the network utilization and the user is instantly informed about the cost of his resource consumption. However, this assumes that the concept of congestion is well established and that unnecessary congestion caused by the party charged is avoided (Crawford 1995). Moreover, congestion is a subjective concept, and some consumers interpret the network as congested, while others do not.

Unfortunately, there are some drawbacks associated with usage based charging. First, the schemes require complex accounting systems, which unfortunately implies heavy loads on the network. Second, the measurement of usage statistics over the entire network also seems to be a problem. Perhaps, governmental policies will have to help coordinating interconnection agreements in order to create a competitive environment. Otherwise, large providers will have a given advantage in competing for customers (MacKie-Manson & Varian 1995). Furthermore, there is traffic, which does not belong to the real content. Pinging and signaling is necessary in order to maintain the set up connection but should not be included in the traffic metering process (Botvich et al. 1997).

Operators' estimates and predictions of future prices for services and applications are regarded confidential. However, rough estimates show that services without guarantees seem to generate the highest gross margins[9] (e.g. the UBR class in ATM), (Botvich et al. 1997).

The QoS Aspect According to Charging

There will also arise questions concerning the interpretation of QoS in 3G systems. In GSM networks, QoS is handled in line with the so-called raw QoS approach where the service behavior varies between a minimum and a maximum level. Services in UMTS will require a consistent and stable average QoS guarantee, with few and infrequent degradations in the service characteristics. This will probably ask for a more advanced interpretation of QoS compared to in GSM (Philippopoulos et al. 1999).

The customer will be able to renegotiate quality agreements, i.e. a customer can disconnect the service if not satisfied or the resource management could initiate QoS renegotiations, as the service level becomes too low for the specific customer. However, stating QoS guarantees turns out to be hard. One reason is that "mobile QoS" consists of several components (a fixed network component, an air-interface component and a handover-related component). The aim however is to offer a radio spectrum with several service classes reminding of those proposed for ATM networks[10] (Morris & Pronk 1999).

Lindberger (1999) investigates how to divide traffic into different traffic classes, depending on different QoS demands. He concludes that only two types of service classes are needed. The traffic will basically be of two types (elastic and stream traffic), where the transfer of

[9] Gross margin: the differential between the production cost and the session revenue expressed as a percentage of the session revenue.
[10] DBR, UBR, SBR and ABR

files of data is an example of elastic traffic and video and voice are examples of stream traffic.

Volume based tariffs irrespective of traffic type is suggested. Stream traffic is offered small buffers that are prioritized in relation to other buffers. The main contribution of the paper is that a satisfying QoS for both traffic types can be combined with a good utilization of the link. This can serve as a good example of how to carry traffic in a multi-service network. It also says that different priority classes for partly the same type of services is ineffective, if the proportions of traffic volumes in these classes is unknown. In the INDEX (The Internet Demand Experiment) project there was an empirical real-world trial of quality-differentiated network services, which provided Internet access over ISDN lines at the Berkeley campus community (Altmann et al. 1999a). Six different variable symmetric bandwidths were offered (i.e. 8, 16, 32, 64, 96, and 128 Kbps) and charged for per minute in relation to the chosen capacity. Experiments with variable asymmetric bandwidth and charging for volume were also performed. The results showed that a significant proportion of the users do have an exact idea of how much they are willing to spend for Internet services, given a certain level of QoS. Further aspects associated with the area of QoS are discussed in section 6.

3. Actors

3.1 A Changing Value Chain

Traditionally, the unchallenged most important actors in the telecommunication market are the network operators (NO). They collect about 90 % of the market revenues, thanks to incomes from voice (UMTS Forum 2000).

New end-user demands and the advancing technology will probably change this market structure. A general assumption is that the role of content providers will grow with the introduction of UMTS. Service providers will try to position themselves in the value chain and in return for billing and customer support receive a percentage of the revenue. Early experiences from i-mode (see section 5.1.2) however highlight the difficulties for content providers to generate revenue. Since content is predicted to cover much of the initial cost of the UMTS infrastructure, this ought to change. The merger of AOL and Time Warner and the emergence of Internet banks like Citibank however confirm the growing interest for the opportunities within the market for content (www.citibank.com).

3.2 Network Operators

So far, the task of the NO has primarily been to enable transfer of voice and data in a market characterized by incomplete competition. The question arises, whether mobile operators have sufficient innovative ability to continue to stay competitive. This challenge has become a question of service management, more than just network management, since the revenue stream will be closely linked to quality of service and the availability of applications (Jagau 2000). Actually, customers are just “one click” away from competitors, which confirms the emergence of loyalty agreements.

There are basically two ways network services could be delivered. First, the facility-based operator (e.g. Telia or Telenor) trades network capacity (wholesale) to ISPs that does not own any network infrastructure. Second, the operator could act as a retailer by selling services to end customers directly. The ISP also acts as a retailer when the service is "resold" to the customer (Ericsson & Persson 2000).

Today license holders do not have the skills and expertise to offer the full range of mobile commerce, entertainment, banking, shopping, information and other possible services that will be available in 3G networks.

The appearance of so-called Mobile Virtual Operators (MVOs), confirms the attractiveness of the business related to the retail business. Other actors in the mobile market are virtual Internet service providers and portals.

3.2.1 Mobile Virtual Operator

A mobile virtual operator (MVO) is a firm without infrastructure, but which issues its own SIM card. The MVOs are focusing on providing software and content instead of delivering access to the local loop, the radio spectrum or the global Internet. The Norwegian Sense Communication and Virgin Mobile from UK are typical examples of MVOs.

The principal question is whether facility-based operators should let newcomers like MVOs enter their market. They can actually increase the supply of content and trigger an increase in traffic volume, i.e. acting as each other's complements. For instance, Virgin Mobile can, as a WAP service provider, create higher demand than what One-2-One (the owner of the infrastructure), with whom they have a joint venture, can do alone.

The NO faces the risk of loosing customers to the MVO and that the enabled services become more of a substitute than a complement to the facility-based operator's services (Foros & Hansen 2000).

3.2.2 Virtual Internet Service Providers and Portals

Just like the MVOs, the virtual service provider rents capacity in the underlying infrastructure. The difference between the mobile and the ISP market tends to be the few spectrum licenses, which constrains the possible number of MVOs. Market prices for wholesale services determine if an actor becomes a virtual or a network owning ISP.

A portal is a start page for a specific segment of users, which could be provided via an MVO. The customers see a portal as a place where services are available (Foros & Hansen 2000).

3.2.3 Internet Backbone Provider

The Internet backbone provider is at the top of the Internet hierarchy. The service includes guaranteed access to the core routing structure, which is driven cooperatively by a few core IBPs. Recently, they have tried to integrate vertically with the market for Internet accesses, searching for competitive advantages (Foros & Hansen 2000).

3.3 Conclusions

Traditionally, the network operator has been the receiver of most part of the revenue stream from the customers (90%). The introduction of UMTS implies a more important role of the content provider. The major explanation is that the facility-based operators will probably not be able to acquire sufficient amounts of skills and expertise to meet new end-user needs. Hence the market value chain will probably look different in a 3G environment compared to today. The appearance of MVOs, virtual ISPs and Portals are typical examples of that.

This change will however require a redefined distribution of the total income. Otherwise, the attractiveness of the market for mobile communication will fade. This would be negative for the UMTS business case. Thus, what is needed is a well-defined framework, which meets the interest of all actors involved.

Problems associated with this are customers' restricted willingness to pay for services and arising conflicts between the branches of tele- and datacommunication. Network operators are probably not too eager to share profits with new actors and at the same time expose themselves to enhanced competition. On the other hand, if an extended number of actors are prepared to bear the burden of investments etc., this could actually lead to profitable scenarios resulting in "win-win" situations.

4. Network Characteristics

Charging structures for mobile services require knowledge of the underlying network architecture. Yet, this is not available. There are lots of theories and some of them tend to dominate over others. Part of the problem is the fact that the traditional branches of telecommunication and computer communication have started to near each other, moving towards a common network structure.

Besides corporate culture conflicts, the utmost important issue regarding this convergence is whether networks should be mainly packet or circuit switched. This provides a reason to point out some differences between the two technologies.

4.1 Packet and Circuit Switched Networks

Packet based networks were originally built for best-effort traffic and lack real-time traffic support. The end-to-end delay could be quite significant, depending on packet reassembly and forwarding at intermediate routers. To preserve real-time traffic properties, packet sizes should be small. This unfortunately leads to a high proportion of overhead, which is not easily compressed over radio links (Yang & Kriaras 2000). Statistical multiplexing however allows effective resource allocation.

IP (Internet Protocol) is the predominating packet-based protocol standard and it is supported on a range of networks with different bandwidths, transports and performances.
In the future there will be features in IP including the ability to reroute connections for long duration transfers as the network detects more efficient routes. This helps to preserve a more reliable time-delivery. Even IP over circuit switched link layers such as ATM (Asynchronous Transfer Mode) can handle traffic of delay critical nature (Marchent et al. 1999).

Circuit switched networks establish a fixed connection between two or more communicating entities. No more than one session at the time is allowed to use the network capacity of that particular connection. Consequently, the resources are not always utilized, which is wasteful. Moreover, the set-up procedure is time consuming.

The immediate benefit of this type of connection is that traffic is delivered to the destination in a timely manner. This is a crucial characteristic for services like voice and videoconferencing.

Despite the benefits associated with the circuit switched technology, IP constitute the most "popular" communication platform at the moment.

4.1.1 "All IP" Networks

Ericsson's vision is an evolution towards "all IP" networks (i.e. that fixed and mobile networks are integrated). This is motivated by the fact that IP is market driven, more than just an issue of technology (Örnung 2000). Another factor supporting the idea of IP networks is the increased attention given to services like IP-telephony.

The PSTN, the mobile networks and the Internet are separated, a structure that is generally believed to change. Nowadays, there is a focus on the establishment of a backbone (presumably IP), common for all kinds of access networks, including UMTS. The idea is to avoid detached networks, with separate but equal equipment (Örnung 2000). The applications will be located on servers directly connected to the backbone. This alternative approach is illustrated in Figure 3.

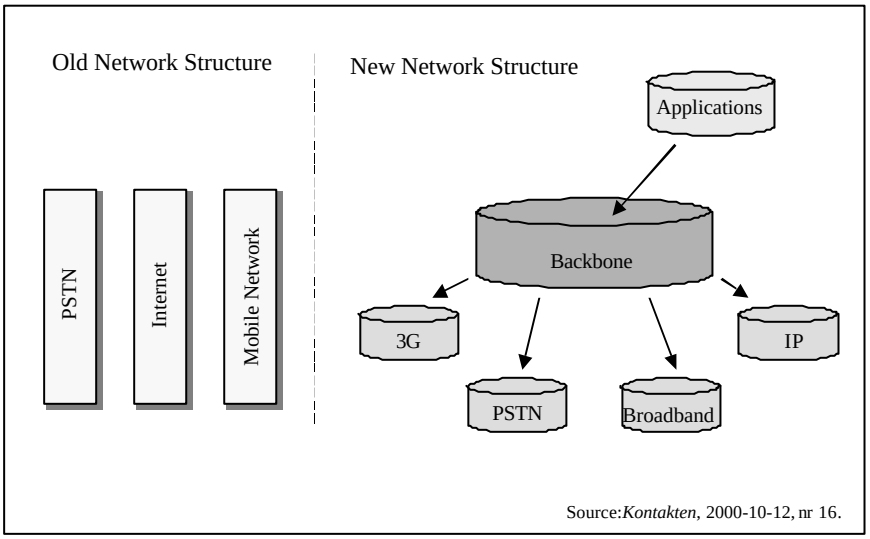

Figure 3. The migration towards a common backbone network structure

With an IP-based backbone, it sounds logical to deploy IP as a platform even for UMTS, which is further studied below.

4.2 UMTS

The term 'all IP based network' is not well defined, but what is meant is to deploy IP as the general platform for all services. The main purpose of this structure according to www.3gip.org is to allow operators to deploy the IP technology to deliver 3rd generation services (www.3gip.org).

These are some benefits mentioned in (Yang & Kriaras 2000) associated with IP in UMTS access networks[11]:

- Operators could easily offer the same services to subscribers accessing through different networks.
- The cost of providing IP transport is continuously sinking.
- An IP-based UMTS core network would mean a smooth interworking with an IP back-bone.
- Capacity enhancements of an IP based transport network are easier and cheaper.

There are also uncertainties related to the deployment of IP in UMTS. It is not solved yet who will be responsible for service control. Will it be the home network, the visiting network or both? Even the administration of user location services is an open question.

Thus, an alternative UMTS system could be a hybrid of packet based and circuit switched systems, which means that the UMTS terminals must be able to handle both of them. By doing so, services which draw from the benefits admitted by circuit switched system do not have to convert to IP and vice versa.

The UTRAN (UMTS Terrestrial Radio Access Network) might be introduced with inter-working to the GSM system. Later, the UTRAN will be connected to a UMTS core network based on either ATM or IP (Yang & Kriaras 2000). The evolution to an IP/ATM core network is illustrated in Figure 4.

[11] This is however not true regarding the radio interface.

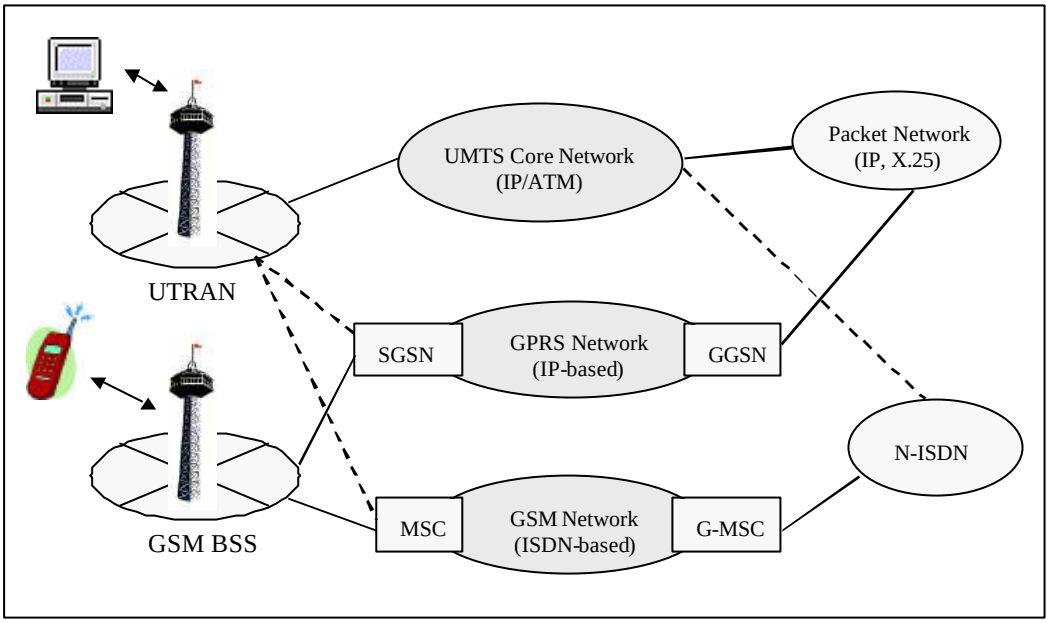

Figure 4. The evolution towards and IP / ATM UMTS core network

This brief description of the UMTS architecture serves as a reference regarding the delivery of applications and QoS. Moreover, the architecture should be kept in mind when discussing future charging schemes.

Next chapter describes what applications seem to be most interesting seen from a 3G perspective.

4.3 Conclusions

The choice of appropriate charging schemes for UMTS services will depend very much of the underlying network structure. Generally, in circuit switched networks, the time connected plays a significant role compared to packet switched networks. On the other hand, in packet switched technologies, the amount of data transmitted is perhaps a better measure when it comes to the consumption of valuable network resources. Besides, different networks serve the needs of some applications better than others.

Circuit switched networks could guarantee the customer a certain network capacity, but unfortunately, this is at the expense of inefficient network operation. IP is the communication platform that obtains the most attention at the moment and the focus is on the establishment of an IP backbone, common for all kinds of access networks. Uncertainties concerning the service control responsibility, demands further investigation. No one would like to pay for a service that is not delivered. Perhaps a hybrid network structure, consisting of both circuit and packet switched characteristics, might be the best solution in the end.

5. Applications

Most surveys related to the area of applications are focused on fixed networks and only a limited number of them are available over wireless systems. Therefore, the mapping of services/applications from fixed to mobile platforms tends to be risky.

There are at least four different characteristics that distinguish mobile and fixed Internet connections from each other (Ericsson et al. 2000):

- Mobility
- 'Always connected'
- Transactional
- Local awareness

Mobility and 'always connected' mean that the user is spared from having to be situated at an Internet access point and does not have to log on to the network every time.

The transactional feature implies that the user can act instantly on information, like buying stocks.

Finally, local awareness implies all terminals to be located at any occasion, enabling positioning related applications. Technology issues affect the evolution of applications in at least two ways. First, network capacity naturally restricts the supply of services and the feasible number of subscribers.

Second, the power of the processor, network capabilities, energy source supply, memory size, screen size etc., will certainly influence the supply of services.

Streaming services, as it seems, require too much capacity, even in the context of UMTS. Consequently, the option of downloading music for instance seems more appealing.

Europolitan recently made the first Swedish release of GPRS. This is the first real field trial of packet-based applications over wireless networks (in Sweden). Hence, results from it will be valuable knowledge in several respects. First, it will indicate which applications are demanded and second, it will show people's reaction towards volume based charging.

5.1 Demand for Future Mobile Applications

People are attracted to different types of applications, though features characterizing mobile communication such as freedom, connection, exchange, closeness and mobility seem to be particularly important to most users. General emphasis is given to applications, which simplify everyday life.

Ericsson Mobile Communications and Ericsson Consumer Lab have performed market research regarding thoughts and demand for non-voice wireless applications. Table 6 illustrates the result from some chosen regions around the world (Consumer lab 2000).

Country	1st	2nd	3rd	4th
France	Tariff indication[12]	Own location	House alarm	Maps
Italy	E-mail	Maps	Information services	Web browsing
Japan	Remote function	Remote function VCR	Buying information and tips	Maps
Nordic	E-mail	Maps	Web browsing	Information services
USA	Own location	Convert to cordless at home/office	House alarm	Remote function

Table 6. The four most interesting non-voice mobile services

The demand for mobile music devices was pretty modest, but still higher than games. Despite a relatively low overall interest for these services, young people seems to be the main target group.

5.1.1 WAP

WAP (Wireless Application Protocol) is a "phone browser technology", i.e. a suit of specifications that defines a protocol for communication between server applications and clients. WAP acts as a catalyst for the mobile Internet. In many respects this has led to a somewhat turbulent situation for many operators. The business models of the mobile Internet are quite similar to those of the fixed Internet, with some important exceptions (Pehrson 2000).

- Portals to the mobile Internet have a more prominent role compared to portals to the fixed Internet.
- The success (or failure) of the mobile Internet is more dependent on classes of service than on mobility enabling technologies.
- The mobile Internet represents a major opportunity for electronic e-commerce.

Wireless operators in possession of large customer bases and gateways between mobile networks and the Internet are well positioned for success in the emerging datacom era. The WAP gateway can be located at the site of the mobile operator, an independent service provider or an enterprise. Each of them provides different advantages. If it is located at the operator, the end user can obtain faster access, customized billing and other specific network features provided by the operator. An enterprise could enable unique end-to-end security (e.g. bank accesses).

The expected penetration of WAP is based on the mobile penetration. By 2004 nearly 95 % of the users with WAP-enabled phones are expected to have active WAP subscriptions overall (Pehrson 2000).

[12] The user is informed about how the charge is calculated.

So far, the time-consuming set-up times over circuit-switched connections are a major impediment. However, this will be eliminated thanks to GPRS (always connected).

5.1.2 Applications over GPRS

Europolitan recently introduced GPRS, which is generally considered to be the first step towards the mobile Internet. GPRS enables the user to be connected to the Internet without paying for the actual time spent online, but for the volume of data transmitted.

GPRS will not support applications of real-time characteristics, but services, which require low QoS and low bandwidth. Services with good potential on the GPRS market will possess the following characteristics (Mörk & Wennerström 2001):

- Useful or entertaining and is not readily available through substitutes.
- Swiftly provided, reliably and efficiently.
- Easy to use with respect to input and navigation.
- Offer a high value compared to the total cost of service for the user.
- Utilize little network capacity, which results in low transport costs.

GPRS will probably be used for qualitative information services, but SMS will continue to be used for the delivery of most quantitative information services (http://gsmworld.com).

GPRS will serve as an indicator of what applications are feasible in a UMTS environment. An important aspect is that services and applications, which are under development today, will be functional, even in that system (Europlitan 1999/2000).

5.1.3 I-mode

I-mode constitutes the best evidence, so far of applications demanded over the mobile Internet. Knowledge about i-mode could be valuable when determining how to approach an effective WCDMA launch strategy.

I-mode is a service offered by the Japanese telecom operator NTT DoCoMo. Since the introduction on February 22 1999, the number of subscribers has grown to incredibly 21,356,000 (http://www.nttdocomo.com).

The volume of voice traffic is predicted to decrease. In 2002, it is predicted to be lower than 55 % and in 2003 voice and data traffic is estimated to be about the same. E-mail is the most popular data service and stands for approximately 40% of the mobile data usage (Scott & Irvine 1999). The supply of other services and applications includes transactional (e.g. mobile banking), information (e.g. news) and entertainment (e.g. karaoke and network games) services (Ericsson et al. 2000).

The networks handle packet-based data and the user is charged per byte. An i-mode subscription costs ¥300/month (i.e. ≈ 26 SEK/month) with an additional charge of ¥0.3/packet[13] (≈ 0.03 SEK/packet) sent (Scott & Irvine 1999).

[13] 128 bytes.

Some services include specific value added, meaning an extra monthly fee paid directly to NTT DoCoMo, who keeps about 9 % of the payment. The rest is transferred to the content providers (Mörk & Wennerström 2000).

The average monthly i-mode bill is ¥10 883 per user (≈ 927 SEK)[14] (Mörk & Wennerström 2000). This figure is related to the personal budget and peoples' ability/wish to spend money on mobile services. It could also be recognized that people's propensity to spend money on mobile usage has increased with the introduction of i-mode since an ordinary voice bill is ¥8 250 (≈ 702 SEK).

To summarize the success story of i-mode, there are two important conclusions to be made, which are useful with GPRS and UMTS at hand.

First, NTT DoCoMo, which is the owner of the network, is the receiver of most revenues. This is, because voice stands for the major part of the revenue stream. The rest comes from service charges, data traffic and billing commissioning. Only a small part of the total revenue is shared with the content providers.

Consequently, content providers create value, but receive little revenue. Evidently, this has to change if UMTS is supposed to be an attractive market opportunity for content providers. Yet, it is unsure how to proceed; all-inclusive rates, hardware subsidization, pre-paid or bundling of services by mobile data providers could be alternative business solutions to consider.

However, performing strategic business plans for UMTS, entirely based on experiences from i-mode might be risky. There are cultural differences between Europe and Japan that cannot be ignored.

5.2 Applications, Resources and Quality of Service

Applications enabled by WAP, i-mode and GPRS are not very resource demanding, compared to real-time traffic such as IP-telephony and videoconferencing. Furthermore, real-time services demand guarantees regarding QoS. However, some applications require more from the network than others do.

Services like IP-telephony are only functional if enough bandwidth is reserved along the entire path between two users, so that the end-to-end delay does not exceed 500ms, which is critical for the human perception. Besides, only insignificant jitter is allowed (Ericsson & Persson 2000).

Videoconferencing requires a minimum bit rate, but will generate frequent peaks in the data load. Moreover, audio, video and data must be synchronized, but there are however no possibility for retransmissions, since the communication takes place in real time. Damaged messages could be repaired by the video codecs.

[14] www.tullverket.se/vaxelkurser/21_31mars2001.htm

Streaming media could be divided into sub-streams thanks to layered coding schemes. Hence this could guarantee the client some basic quality and a higher quality if resources allow it (Engman 1999). This means that the media could be delivered to the customer with varying quality. Some users may be satisfied with that, while others will not.

Services such as ordinary file transfers cannot tolerate losses and delays and jitter in the transfer do not affect the perceived QoS notably.

Below, some applications/services are listed, aiming at describing the variety in QoS requirements. It should be noted that the demand for resources fluctuates over time, which means that the support for QoS could be dynamically allocated (even for a single application).

Type of Application	Delay/Jitter Tolerance	Loss Tolerance	Bandwidth
IP telephony	Low/Low	Low	Hard bw guarantees
Video Conferencing	Low[15]/Low	Low	Hard bw guarantees
Interactive games	Very low/-	Very low	Guarantees
Online banking	Very low/-	Very low	Hard bw guarantees[16]
Adaptive (e.g. ftp)	Moderately/-	Very low	Guarantees

Table 7. Some applications/services with corresponding QoS requirements

Brandt et al. (1999), divides the traffic into four categories. The classes are listed below and they are suggested to the ETSI as possible bearer services for the UMTS. The groups all have unique demands on real-time performance, bandwidth, throughput and availability.

- Background traffic, e.g. e-mails, SMS and downloads from databases.
- Interactive traffic, e.g. web browsing and data requests from remote equipment.
- Real-time streams, e.g. real-time video/audio (always one direction).
- Real-time conversation, e.g. speech-telephony, VOIP (always between human peers).

Since the radio frequencies are scarce, compression techniques like MPEG-4 and H.263 will be very useful in order to decrease the required bandwidth on the underlying network. However, replaying compressed streams obliges significant computing power. A 200 MHz Pentium 64 MB can hardly handle more than a 1.5 Mbps H.263 coded video (Engman 1999). If the streams are uncompressed, less computing power is needed at the expense of increased resource consumption. Here is a collection of applications and their corresponding bandwidth needs.

[15] Broadcast information could though be buffered at the destination, which makes it rather insensitive to delays.
[16] Transactional

Applications	Bit rates (Mbps)
Real time audio streams (CD quality) (Ericsson & Persson 2000)	1.411
Mp3 sound (CD-quality) (Ericsson & Persson 2000)	0.128
Mp3 sound (radio-quality) (Ericsson & Persson 2000)	0.056
Mp3 sound (telephone-quality) (Ericsson & Persson 2000)	0.008
MPEG-1 video coding (Ericsson & Persson 2000)	1.86
MPEG-2 video coding (VHS quality) (Ericsson & Persson 2000)	2
MPEG-2 video (DVD quality) (Engman 1999)	4-6
MPEG-2 video coding (PAL quality) (Ericsson & Persson 2000)	4
Uncompressed PAL (Engman 1999)	216

Table. 8 Some real-time applications and their corresponding bit rates

Obviously, the need for quality support is nothing given. It depends on what the customer has in mind and subjective perceptions about what constitute good quality. Hence, what is meant with service guarantees differs from one service to another. In the next chapter, the concept of QoS will be further investigated.

5.3 Conclusions

The experiences obtained so far considering mobile communication over packet switched networks (WAP excluded) come from early trials of applications and services admitted by i-mode and GPRS. Common for them is that voice is the uttermost demanded service and that e-mail is the most attractive non-voice data service. Another important aspect, especially regarding i-mode, is that the average age of the users is remarkably low. That means that people in their teen ages could be the most potential user group in a future perspective.

There are several lessons to be learned from WAP, i-mode and GPRS that could actually be valuable to bear in mind when introducing UMTS. Here are some of those observations:

- Portals to the mobile Internet have a more prominent role compared to portals to the traditional Internet.
- The success of the mobile Internet is more dependent on classes of service than on mobility enabling technologies.
- The time-consuming set-up times over circuit-switched connections are disturbing.
- Prices incorporate impediments to usage. Besides, forecasting the price for utilizing simple services such as reading the newspaper turns out to be tricky.
- The market as it looks today is not very appealing to content providers, since most of the revenue stream goes to the network operators. This ought to change, since the margins on data services are more tempting (see SMS and GPRS prices). Furthermore, coming UMTS investments are supposed to be financed to a great extent by content.
- The average monthly bill has increased (see i-mode).

It is also crucial to understand that different services have different requirements on the network, even though many of them possess real-time traffic traits. IP telephony requires insignificant delays and jitter and tolerates only minor packet losses. Videoconferencing is

only performable if a minimum constant bit rate is guaranteed. Moreover, audio, video and data must be synchronized. Services such as ordinary file transfers cannot tolerate losses. Delays and jitter in the transfer do however not affect the perceived QoS notably. Hence, charges must take these heterogeneous characteristics into consideration, since what is meant with service guarantee differs from one service/application to another.

6. Quality of Service (QoS)

QoS is a blurred concept, but proves to be very important for many applications and services in a UMTS context. Unfortunately, there are numbers of definitions of QoS, which certainly does not make the situation easier.

According to ISO 8402, quality is interpreted as the "*totality of characteristics of an entity that bear on its ability to satisfy stated and implied needs*" (Kokkonen 2000, p.4). The definition according to ITU-T, E.800 "*The collective effect of service performance which determines the degree of satisfaction of a user of the service*" is perhaps more useful in the context of mobile communication (M3I 2000b, p.9).

Usually, QoS is often confused with what people frequently call service quality. However, there are differences, depending on which view is taken. QoS is more appropriate when studying the communication services[17], while the notion of service quality is referred to as the service delivered in the interaction between a user and a service provider.

What is meant with "service" is another reason for misunderstandings. First, service levels may vary substantially from one service provider to another. Second, two customers might have completely different perceptions of the same service. Consequently, when defining QoS, it is critical that what is actually meant is generally accepted. Hence, the provisioning of an end-to-end QoS is very important in the eyes of the consumer. This is true irrespective of whether the bottleneck is in the radio spectrum or if the traffic is exposed to congestion in any other network. End-to-end QoS is further discussed in section 6.1.

6.1 End-to-end QoS

The mobile traffic passes several networks when it is transferred from one end-user to another. The obvious question is how to allocate resources along the entire path. End-users ought to be guaranteed some objective service level dedicated to a specific service or application. The obvious question though, is how the resource allocation is supposed to be arranged. The task is challenging, since traffic passes several networks on its way from one end-user to another (see section 4). Even if traffic passes national boarders, service levels must be preserved.

ATM was originally designed in order to deliver real-time services, supporting well-defined traffic classes (see background). The trend is however in favor of IP (see section 4). Some type of connection-oriented IP must be designed, in order to resemble the connection-

[17] QoS is explained by technology and capacity issues and is charged for according to well-established contracts (SLAs).

oriented ATM (Brandt et al. 1999). How could then an end-to-end QoS be offered in practice? Vrins (2000) mentions three main alternatives:

- Select the path based on SLAs[18] (Service Level Agreements)
- Pre allocate resources and assign individual bit streams to exclusive[19] paths.
- Adapt paths to network utilization.

The first two are intimately related to the notion of traffic engineering, while the last one is related to IP routing, which is the only manageable alternative currently.

The question though remains where to physically locate the measurement of throughput and latency etc. It is also unsure what should be measured or controlled (packet loss, collect byte counters etc.) and when to measure (real-time or any other time scale).

Hence, what is actually needed in the future is a worldwide connectivity, presumable IP, including excellent coverage and well-established QoS contracts. There are many ways this could be realized. In case of global management, the service provider could build a network by itself, create joint ventures, join a consortium or purchase services from other service providers (Vrins 2000).

Interconnections could be established in different ways. Either, they are constructed by mutual agreements, which is characterized by lengthy, tiresome processes. There must be agreements about SLAs and technical design topics. Clearinghouses (see section 9) aim at finding the common denominator service, making SLAs hard to achieve. Finally, standardized bilateral agreements are comparatively easy to implement. A possible drawback could be a limited supply of service classes, required by certain applications (Vrins 2000).

Interconnection agreements must describe service features, the technical design aiming at fulfilling the negotiated services and what business model to use, including pricing, accounting and settlement issues.

The most well known proposals, which enable QoS between end-users, are IntServ (Integrated Services) and DiffServ (Differentiated Services). They are described in section 6.1.1. and 6.1.2 respectively.

6.1.1 IntServ

Integrated Services (IntServ) offers QoS, since resource requirements associated with a certain application is reserved in advance along the entire path across the network(s). There are two main traffic classes, i.e. *controlled load* and *guaranteed service* (Ericsson & Persson 2000). The first enables relative priority and the second treats traffic in line with beforehand given parameters, like delay and transfer rates. Resources are reserved along the path by protocols like RSVP (Resource Reservation Protocol). This is illustrated in Figure 5.

[18] SLAs are agreements on the management level.
[19] Both different and disjoint.

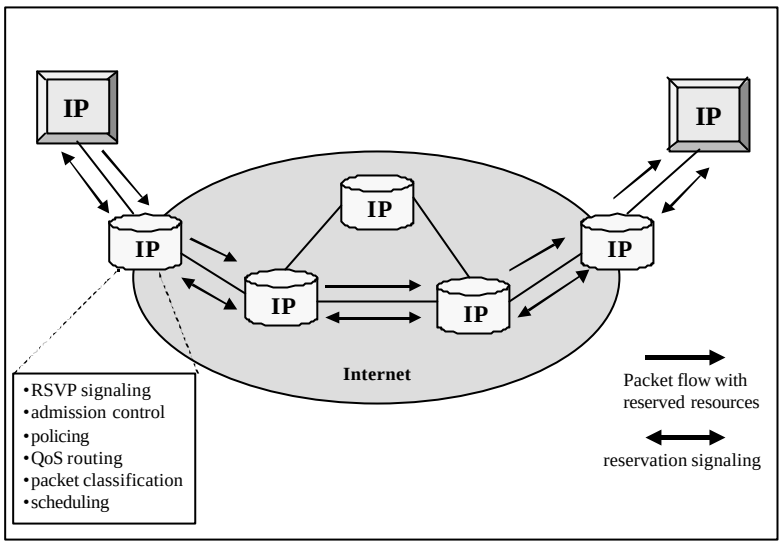

Figure 5. The reservation of network resources in IntServ

IntServ is not suitable for high volume IP traffic, since it requires lots of connection states in the routers. Instead this type of architecture is more appropriate for low volume multicast applications and on demand reservations for wideband packet flows (Brandt et al. 1999). The fact that IntServ complicates installations has opened up for alternative ways to provide quality of service.

6.1.2 DiffServ

Differentiated Services (DiffServ) provides QoS through traffic classes with inherent priority levels. Priority bits are set in the packet headers serving as routing and policing information according to Service Level Agreements (SLAs) at the network edge. Since the interior routers schedule packets depending on their relative priority bits only, DiffServ is more scalable than IntServ (Brandt et al. 1999).

It could be more practical to centralize the resource-handling functionality into a separate server called a bandwidth broker. Its purpose is to allocate resources dynamically over the network. Bandwidth brokers in different administrative domains communicate with each other in order to reserve resources along the entire path. Figure 6 shows features inherent in DiffServ.

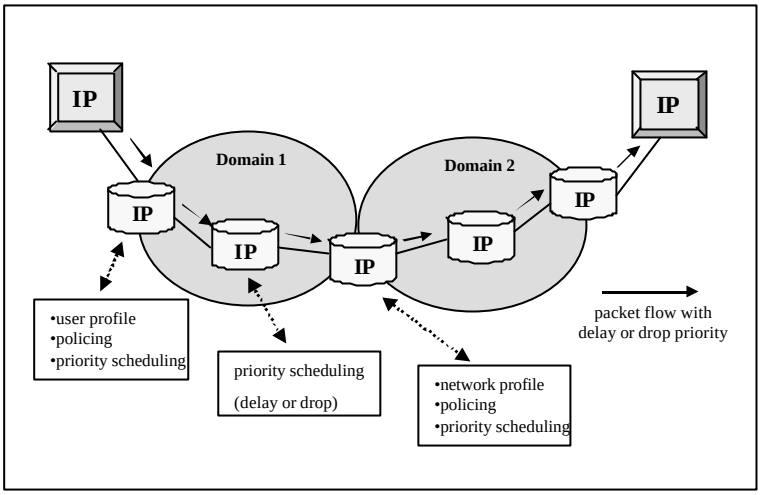

Figure 6. The reservation of network resources in DiffServ

6.1.3 End-to-end QoS in UMTS

The discussion about QoS has so far remained conceptual. However, what seems to be of particular interest is how to provide end-to-end QoS in UMTS networks. Unfortunately there is no single answer.

There are four traffic classes defined for UMTS, i.e. background traffic, interactive traffic, real-time streams and real-time conversational (described briefly in section 5.2)[20]. Traffic, as it seems, will not be statically fixed to a certain traffic class. Depending on the temporary congestion, traffic will probably switch dynamically between different classes as the supply of resources fluctuates.

Before entering, an IP backbone (or a circuit switched equivalent), traffic must be mapped against some other service class offering the same quality as in the access network. In addition, the number of classes may differ from one network to another, making the situation even more complicated. This is a critical issue if an end-to-end QoS is supposed to be obtained.

The UTRAN is believed to be the weakest link in the UMTS system due to the scarcity of radio spectrum and will therefore constitute a bottleneck (Ouchterlony & Molin 2001). However any form of quality differentiation in the radio interface will constitute a significant competitive advantage in the future.

6.2 Conclusions

Many applications require certain levels of QoS. The most important property undoubtedly is the establishment of an end-to-end service quality. Even though the radio interface constitute the weakest link and also the most expensive one, the user will be charged for all net-

[20] Further information about UMTS service classes could be found in (Ouchterlony & Molin 2001).

work resources consumed from one point to another. How this is realized, is still unsure and the work within this area is still in its infancy. However there are some good proposals including for example clearinghouse functions and SLAs. The knowledge about users' perceptions of quality of service and the way it is guaranteed is necessary when deciding upon a suitable charging scheme.

7. Demand for Service

Finding the market demand curve for mobile services turns out to be a difficult task. The real world is very complex and dynamic, which makes assumptions and results from such studies unreliable. The progress of the underlying technology evolution proceeds continuously, changing the market rules from one day to another. Since customers' demands could change rapidly, precisely drawn demand curves will be superfluous.

The supply of qualitative surveys within the area is insufficient, to say the least. Besides, most information within the area is classified, due to its close linkage to the business profitability.

However, this section will investigate how to approach the estimation of market demand curves and how this applies to the area of mobile communication services. General factors that tend to influence the demand for mobile services are also discussed.

Demand patterns are mainly determined by the overall IT literacy, which fluctuates widely between demographic groups. The reason for usage varies and each user strives to attain some kind of individual satisfaction, which certainly depends on age, personal budget, price-sensitivity and profession. Hence, an investigation of parameters associated with the customer's choice seems justified, before getting on with the challenging task of drawing demand curves.

7.1 The Consumer's Choice

7.1.1 Values

The consumer's choice is the result of subjectively perceived user values. Thus, the market must be segmented if the needs of specific user groups or market segments are to be found. Of course, the demand for services will strongly depend on whether the user subscribes to them as a residential or a business customer. Since more and more people start working at home, traditional methods of segmentation will however change.

The perceived value is primarily related to the actual services or applications demanded, but also to all other steps through the value chain. The service must fulfill the basic purpose according to the user's needs (e.g. the appropriate information in a timely manner). Here, aspects such as service features, the ease of use, price and overall performance are utterly important. Other features such a s the sales process must be appealing, considering accessibility, reliability and knowledge. Moreover, activities that do not have anything to do with the actual service, as billing and customer helpdesks also tend to affect the perceived value. In conclusion the total service management system matters to customers.

7.1.2 Age

People in the same age often have many things in common, both regarding their social lives and when it comes to their marginal propensity to spend money on communication services. In spite of the relatively high costs for mobile usage, the market seems to be driven by the demand of young people. For example, teenage girls are said to be one reason to why i-mode has become so successful (Tamm 2000). They use the phone, not only as a communication device, but also as symbols that strengthens the personal image. In general, many teenagers are so called early adopters, which means that they are among the first to try new interesting services/applications. For example, applications like mobile games and downloaded music seem to attract youths more than adults.

Adults seem particularly interested in services dedicated for a certain purpose. Applications which strengthens the feeling of control, give the opportunity to remotely control things (security purposes) or simply enables practicality, are typical examples of services that appeal to adults (Ericsson et al. 2000). Besides, grown ups can more easily control the money supply and the personal budget compared to young people, due to the fixed monthly incomes. Budget constraints are further discussed in the next section.

7.1.3 Budget Constraints

When it comes to people's spending on communication services, the budget constraint plays an important role. This section will treat the budget constraint as a relative measure, aiming to illustrate the logic behind people's choices when facing more than one alternative.

Customers' incomes are supposed to cover the cost of many needs. Therefore only a limited part can be spent on these types of activities. In order to illustrate individuals' preferences, one can draw indifference curves. They are constructed by connecting bundles of commodities, (not necessarily physical), between which the customer is indifferent (Katz & Rosen 1998). The slope of such a curve tells about *the marginal rate of substitution* (MRS), i.e. the rate at which the customer is willing to trade one thing (applications, services, methods of communications etc.) for another. Indifference curves can be drawn through any point, resulting in a collection called the indifference map. Points far away to the northeast signify relatively higher customer utility. The shape of the curves is determined by a personal utility function, given a specific service or application in mind.

Actually, these types of comparisons are necessary for the success of the UMTS business case. Imagine that the marginal rate of substitution concerns the relative utility of using EDGE compared to UMTS. The willingness to switch from EDGE to UMTS is signified by the slope of the indifference curves between the two service barriers. Budget constrains then determines the optimal choice of consumption. The way the budget restricts the feasible set of achievable combinations of consumption is illustrated in Figure 7.

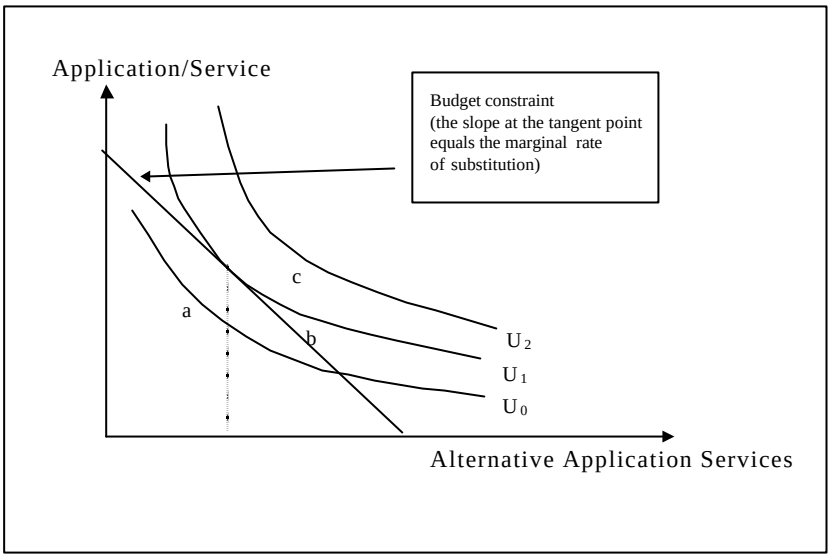

Figure 7. The optimal choice of consumption under a budget constraint

The effect of a change in the relative price twists the budget restraint or shifts it either to the right or to the left, depending on if prices rise or fall. The equilibrium point is found where the budget line is tangent with an indifference curve. At this point the marginal rate of substitution equals the relative prices of the services, that is $MRS = \frac{P_x}{P_y}$.

To make things even more complicated, the outcome of a change in prices is rather ambiguous, since people's utility function could look very different. Hence, understanding the market behavior proves to be extremely difficult and what is needed is a feedback mechanism, which constantly signifies people's behavior. Tariffs must be reconstructed in line with feedback information from the market. The relationship between reserved network resources and actually used resources is a typical example of information, which is useful for both VASP and operators in their aims at optimizing the usage.

Finally, when the individual demand curves are derived (or at least approximated), whether they illustrate specific services or service groups, they can in turn be used in order to derive the market demand. The variety in the shapes of such curves will be great.

7.2 An Introduction to Price Elasticity of Demand

The demand curve embodies important information about people's reactions to price changes. The change in demand caused by a change in price is called the price elasticity of demand (Katz & Rosen 1998). This is written as:

$$e = \frac{\%\Delta Q}{\%\Delta P}$$

If the price elasticity of demand is significant, that implies that a small change in price leads to a relatively large change in demand for a service and vice versa.

Price elasticity is useful when studying how a change in price will affect the total expenditure on the service. As can be noticed the total expenditure may paradoxically be influenced in a positive direction (people spend more money on the service) irrespectively of whether the price goes up or down. Figure 8 illustrates a decrease in the price for the alternative application/service, which could be noticed by the outward twisted budget constraint. What happens to the usage after the price decrease is unsure, since it depends on the level of utility, which differs from one user to another. To be specific, point e3 is further to the left than e1 in the figure (i.e. the usage of the alternative application decreases despite the lower price, whereas at point e2 and e4 the opposite is true).

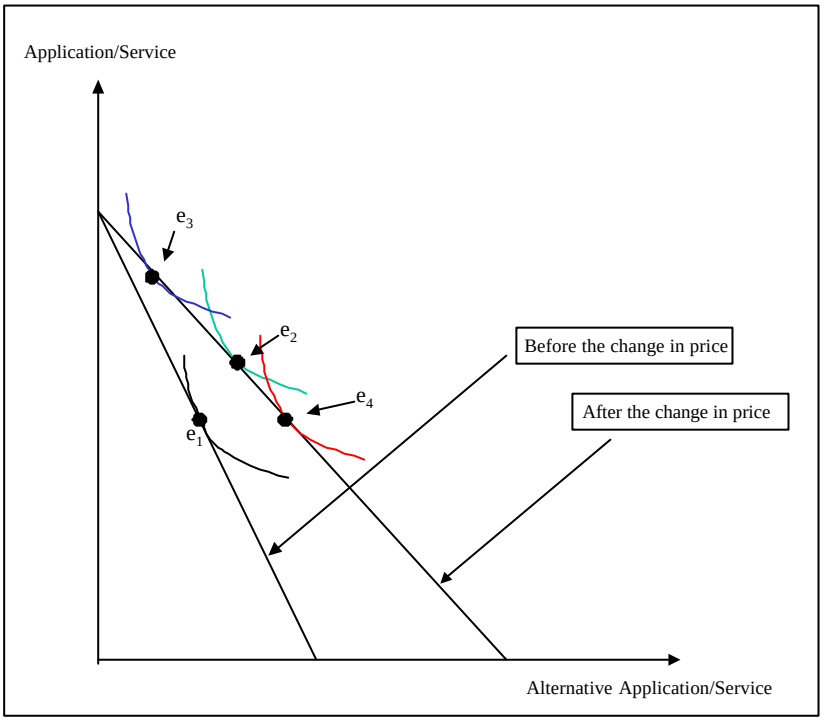

Figure 8. The money spent on services after a change in relative prices

An increase in usage despite an increase in the price may sound very odd, but the way people act is not always rational[21]. Understanding the human psychology is a most challenging task.

Since determining the demand curve turns out to be difficult, it also includes the estimation of price elasticity, by definition. The expression for price elasticity could be written as

[21] Another example of the unpredictability of peoples' demand for services is the sudden and incredible demand for Tamagutchis a few years ago.

$e = \frac{\Delta Q}{\Delta P} * \frac{P}{Q}$[22]. By using calculus, the discrete change $\frac{\Delta Q}{\Delta P}$ could be replaced with $\frac{dQ}{dP}$. The expression then looks like, $e = \frac{dQ}{dP} * \frac{P}{Q}$.

The benefits associated with the latter are when there is an estimated demand curve. Assume that the market demand is reflected by a function of the form $Q = Ae^{-bp}$[23], then the price elasticity of demand equals -βp. Even if the function just mentioned is derived, there are some important factors that tend to affect the price elasticity specifically. Here are three examples (Katz & Rosen 1998):

1. The price of close substitute for a commodity tends to make demand more elastic.
2. The elasticity depends on the commodity's share of the consumer's budget.
3. The elasticity depends upon the time frame of the analysis.

However, even prices of *other services* could affect the usage of a service. Here, a more appropriate measure seems to be the cross-price elasticity, which is defined as the change in demand caused by a change in the price for another service.

$$e_{xy} = \frac{\%\Delta Q_x}{\%\Delta P_y}$$

Here, what seems to be of interest is the existence of complements or substitutes to the service or application in question. Further information about these factors is found in Appendix A.

The question then arises, which services and applications are to be recognized as complements or substitutes to each other. It is possible that a lower price for voice could actually cause a dampened demand for text messaging. This would imply that these services are substitutes. On the other hand, if the price for text messages increases, that could also dampen the demand for phone calls. The reason would be that text messages often cause a phone call, i.e. an increased price for messages affects the demand for voice negatively. In this case, text messages and voice seems to be complements. There are probably lots of other real world examples of this kind, which must be analyzed, in order to get a feeling of what a change in price might lead to.

7.2.1 Residential Users

The price elasticity of demand associated with residential users seems to decrease. The explosive increase of messaging like SMS constitutes good evidence for that. The trend that people do not consider prices very much when making phone calls, seems to include mobile telephone usage as well. Today, people seem to be more dependent on mobile phones, compared to fixed telephones than was the case a few years ago. Overall, people use their phones more frequently and the average holding times are longer (Europolitan 1999/2000). People

[22] Q equals the quantity and P the price.
[23] This type of function must be verified by empirical data.

become more and more dependent on their mobile devices, which probably make them less sensitive to price changes.

7.2.2 Business Users

Business users differ from residential users in the extent to which they are personally responsible for paying the phone bill. As an effect, business users will probably not think too much about the time spent online. A qualified guess is therefore that business demand is more inelastic compared to residential users.

Business users make phone calls that are related to the work and indirectly to the firm's profitability, to the physical location, type of activity and to what extent there are international relationships. This ought to be compared to residential users, who first of all use communication services, for optimizing personal utility.

Ericsson & Persson (2000) refers to surveys regarding price elasticity of demand of fixed telephony, performed during the 70s and 80s. These show that business demand is more inelastic. However, such knowledge must be treated with caution, since mobile traffic in several respects differs from fixed traffic. Additionally, the market conditions have changed a lot during 20-30 years both regarding regulations and opportunities enabled thanks to high-speed networks.

7.3 Conclusions

Predicting people's demand curves seem very hard, since there are so many influencing factors. Some of them are the overall IT literacy, age, personal budget (price-sensitivity), profession but even the total service management system.

What could be done, though, is to determine in what extent services/applications are substitutes or complements to each other and try to figure out the outcomes of individual price changes. Moreover, approximated demand curves could be used when estimating the price elasticity of demand for each case.

However, there is a contradiction when it comes to prices and demands for mobile services. People seem to spend more money on mobile services today, compared to a few years ago. On the other hand, lots of free services are taken for granted over fixed networks. So far little evidence proves that people are prepared to pay a premium price for UMTS applications that are already available over fixed networks. Hence, caution must be taken before something like that is taken as given.

8. Traffic

The aim of this section is to describe some factors behind the traffic environment.
Traffic scenarios are important mainly for three reasons. These are:

- Network planning
- Performance analysis

- Network monitoring

However, forecasting the demand for traffic proves to be deceptive, since consumer behavior is partly unpredictable (see above). However, what is demanded is a function or a model that replicates the traffic environment.

8.1 Traffic Characteristics

A function, A_{t*} is proposed to divide the issue into manageable parts, which have certain advantages. The function contains information about the arrival rate of requests for different applications, holding times, effective bandwidths (described later), reference time factors, penetration rates and the number of customers. The relationship is illustrated in the expression below.

$$A_{t^*} = \sum_{c} \left[\sum_{a} l_{c,a} \left\{ \sum_{b} h_{a,b,t^*} * B_{a,b,t^*} \right\} * R_{c,a} * r_{c,a} \right] * N_c$$

A_{t*}	the traffic load resulting from a group of traffic flows
$l_{c,a}$	the arrival intensity of application *a* for customer segment *c*
h_{a,b,t^*}	the holding time of traffic flow *t**, being part of application *a*
$B_{a,b,t*}$	an effective traffic load measure of traffic flow *t**, being part of application *a*
$R_{c,a}$	a factor describing the reference level compared to an average level for usage of an application by customer group *c*
$r_{c,a}$	the penetration ratio of application *a* for customer segment *c*
N_c	the number of sources in customer group *c*

8.1.1 Arrival Intensity

Modeling the arrival rates of sessions over traditional telecommunication channels often assumes Poisson like distributions. Such assumptions seem to be inappropriate for modern Internet traffic (Fenger 2000). Thus, traffic generated over the UMTS may be quite hard to describe mathematically, since it probably includes features, which are partly characterized by traditional teletraffic (e.g. voice) and partly something else (e.g. data services).

Generally the arrival rates will grow significantly during the years to come. New subscribers will cause parts of that growth. People also tend to use the mobile device more, since new services and applications invoke people's needs and claims.

In Yankee Group (2000) some assumptions and predictions are made, regarding the presumed mobile phone usage 2005 in Europe.

- 95 % of all mobile phone users will be using some form of messaging or e-mail service.
- 68 % of all users will be using information services.
- 14 % of business users will be using mobile communication technology for LAN access[24].
- Just over 3 % of all users will be using video services.
- Telemetry applications will contribute with just over 1 % of total service revenue (voice plus data).

Of course, network operators as well as infrastructure developers strive to attain good knowledge about how to model wireless traffic. Ericsson & Persson (2000) describes a study over requests to three web proxies, performed during 1999 in order to find out what applications are used within the HTTP protocol. Surveys covering fixed network traffic could serve as rather good hints about presumable wireless traffic scenarios[25]. The study presents facts about document sizes and request frequencies. It is noteworthy that video stands for less than one per thousand of the total amount of requests, but still generates eight percent of the total traffic (mean size 1.2 MB).

	HTML	Image	Formatted	Audio	Video	Other	Total
% of requests	20.6	73	0.54	0.30	0.078	5.5	100
% of Bytes	19.6	44.7	11.1	3.7	8.3	12.6	100
Mean size (KB)	10.9	7.0	236	142	1224	26.3	11.5
Median size (KB)	4.7-5.0	2.2-3.2	6-10	3.6-16	370-440	3.2-24	2.5-3.5

Table 9. Typical applications executed over the HTTP protocol

8.1.2 Holding Times

Holding time will constitute an informative measure even in UMTS. In 'all IP' networks traffic is multiplexed and users could stay online all day (without paying). Consequently, one could argue that the value of measuring holding times has declined. Still, there are reasons why holding times are informative. For example there are component such as voice coders that are locked up for a certain time, during a session (irrespective of circuit or packet switched network platforms), which makes the measurement of holding times motivated.

Furthermore, some applications will still ask for circuit switched connections even in the future. Streaming applications are good examples, since they demand excellent QoS and cost effectiveness[26]. The average play (holding) time for such an application will probably not be very long, due to factors like poor screen sizes etc. Notwithstanding, the amounts of data could be quite significant, which is recognized above.

[24] In Sweden, 30% of the total customer base belongs to the business segment.
[25] Resources restrict what type of services are performed over wireless networks. Moreover, long waiting times are not compatible with UMTS applications.
[26] Streaming applications will probably become quite expensive over circuit switched connections.

When it comes to packet-based traffic, there is a need to allocate just enough resources in order to manage the transfer. Here, the concept of effective bandwidths could be used as a conceptual way of approaching the issue at hand.

8.1.3 Effective Bandwidth

In Kelly (1995) a charging scheme for bursty high priority connections based on the effective bandwidth is proposed. The scheme involves trade-offs between the user's uncertainty about traffic manners and the network's ability to statistically multiplex connections efficiently.

As long as the sum of effective bandwidths is less than a certain level, the resource can deliver a performance guarantee, which will be crucial for coming multimedia applications. Depending on the load produced by a source within a certain time interval, the constraints regarding effective bandwidths varies accordingly. Some sources have strict delay requirements and are hence prioritized. The effect is that sources belonging to some traffic class are treated similarly.

All sources have some peak rate *h* and mean rate *m*. The effective bandwidth of such an application is increasing and concave in m for a fixed h, while for fixed m it is increasing and convex in h. If h is small with respect to the link capacity, the effective bandwidth approaches m. In contrast, if h is large compared to the overall capacity, the effective bandwidth will get close to the peak rate of the source. How effective bandwidth could be used for charging is discussed in section 9.1.2.

Gibbens (1996) illustrates the way the effective bandwidth is dependent on the packet size and the time of arrival, by computing the effective bandwidth surface from the Star Wars MPEG-1 video trace. Graphs over such surfaces show that the uncertainty about the quantity of data arriving during a particularly small interval of time is reflected in rapidly increasing effective bandwidths.

8.1.4 Reference Time Factor

In Iversen et al. (2000) it is discovered that the inter arrival rate of fixed Internet traffic is unevenly spread during a day. Obviously there is a peak in the usage between 20-23 o'clock in the evening, which is not strange, since that is the time when people are free to execute their bank errands, information research, etc.

The UMTS usage could very likely look different from that of the fixed Internet. UMTS offers services and applications, aiming at simplifying the everyday life. People may for example execute bank errands during their spare time, (e.g. at lunch or on the bus). Further, trading results and stock information is typical information that best serves its purpose during daytime. An appropriate assumption could therefore be that usage would be more evenly distributed during the day, compared to fixed Internet arrival rates.

8.1.5 Penetration and the Number of Sources

In UMTS Forum (2000) there is a forecast of the worldwide demand by region for some selected services from 2005 to 2010. Services are classified into different business models, which are customized infotainment (mobile portal), mobile messaging services (mobile spe-

cialized services) and mobile Internet/Extranet access (Mobile ISP). According to the forecasts, infotainment stands for the most progressive increase in the number of subscribers and revenues in Europe. Table 10 demonstrates a few predictions of the numbers of subscriptions in Europe by 2005 and 2010.

Subscriptions in millions	2005	2010
Customized infotainment	22.6	100.2
Mobile Internet/Extranet	7.4	69.7
Multimedia messaging services	10.1	43.9

Table. 10 Predicted numbers of subscriptions 2005 and 2010

This prediction does not say much about the Swedish market, but the percentage increases will probably be about the same. It should be recognized however that the mobile phone penetration in Sweden is about 65 %. Hence the number of additional users will be restricted (Mörk & Wennerström 2001).

It is recognized that the penetration of these services will not be very high before 2005. Considering the mobile Internet/Extranet access, it is said to be slow, since it is targeted mostly for the fixed networks. The breakthrough over wireless links will probably wait for relatively high transmission rates and better QoS (UMTS Forum 2000). Lack in security will further delay a major increase.

8.2 Conclusions

Since estimating future mobile traffic scenarios turns out to be a very complex research problem, one has to study separate parameters, which all contribute with useful information. As it seems, the arrival rate will grow much because of a major increase in the information requests. The demand for video services is believed to be quite modest, but the traffic generated per request will be significant. The occurrence of network peak loads (congestion) may differ from those of fixed networks, since much of the usage serves the purpose of simplifying everyday life. Thus, perhaps the overall load will be more evenly distributed during the day.

However, in order to optimize resource consumption, there are methods, which prohibits unnecessary bandwidth allocation. Estimating the so-called effective bandwidth is one of them. The main idea is to foster the consumer to better estimate the needed bandwidth. Otherwise, he is subjected to a higher fee. The concept of effective bandwidth seems to be a bit too abstract for the average user. It however serves as a good conceptual example of how bandwidth could be allocated.

The penetration will also increase during the next ten years. For example, the number of Internet/Extranet subscriptions is forecasted to be eight times as many in 2010 compared to in 2005. The number of infotainment and multimedia messaging subscriptions is predicted to be approximately five and four times as many in 2010 as in 2005 respectively. The growth will probably not be very rapid in the beginning, partly because of lacking transmission rates, QoS guarantees and security issues.

9. Charging

Often there is a mismatch between the technical and the business approach to charging. The assumed increase in the distribution of content in 3G networks will ask for charging bases which take the individual application or service into account. Thus, the way these activities are charged for will vary from case to case. However, people in general do appreciate charging schemes that are easy to understand and that well suits their personal habits.

Perhaps, the problem is not essentially how to measure actual usage, even though it will require the storage and computation of vast amounts of data. Instead the main problem might be to decide upon what to charge for. There are certainly parameters, which are more important than others, but revealing them in this early stage is not allowed, due to confidentiality policies.

9.1 A Charging Framework

What is needed is a charging framework, which is independent of particular business models or purposes of the single actor.

In the future, the value chain in the area of telecom and datacom will be extended. Hence the number of chargeable events will presumably increase. A possible scenario is that the customer loyalty will decrease and that the switching between different operators, content providers and so on, will become much more frequent compared to today.

This in turn will ask for some kind of generic scheme, that is able to charge the customer in real-time. However, much of the functionality of the network is still needed. For example, the customer should still be given the opportunity to trace faulty transactions and control accounting. At the same time the customer must be aware of the cost for his usage. Here, it is up to the service provider to negotiate debt and credit policy depending on the subscriber's creditworthiness.

Ericsson is one of the leading vendors, regarding charging in the B2C interface. (i.e. towards the end user).

9.1.1 A Real-time Charging Mechanism

Real-time charging can be realized by a combination of the charging control function and an on-line communication between the serving network element and the charging function.

The end user requests a chargeable event, invoking the serving element to send charging parameters to the charging function via an on-line protocol. In the charging function some type of credit check or pre-reservation is performed. Meanwhile, the serving element is placed on hold. At last, the charging element is sent from the charging function back to the serving element and the request is either performed or stopped.

At the moment, the Advice of Charge (AoC) by 3GPP (www.3GPP.org) is one of few standards, for real-time end-user charging.

9.1.2 The Charging Function

The number of charging functions is large. However, there are some general characteristics that are valuable to bear in mind when searching for a well functioning charging scheme.

First, if there is a tariff that is based on the mean rate of a traffic flow, there will be no incentives for the user to limit the max rate. Similarly, if the tariff is based on the max rate, there will be no incentive to limit the mean rate[27]. However, charges based on effective bandwidths do in fact take both these restrictions into account and determines the ideal capacity requirement for a given service, i.e. the customer only pays for the prediction of his own usage.

Kelly (1996) has proposed a charging scheme that is based on usage according to:

$$C = a*T + b*V + c$$

where T = time, V = volume and a, b, c are per-time, per-volume and fixed charges respectively.

The parameters a and b are determined according to Figure 9 below (i.e. where C/T is tangent to the effective bandwidth).

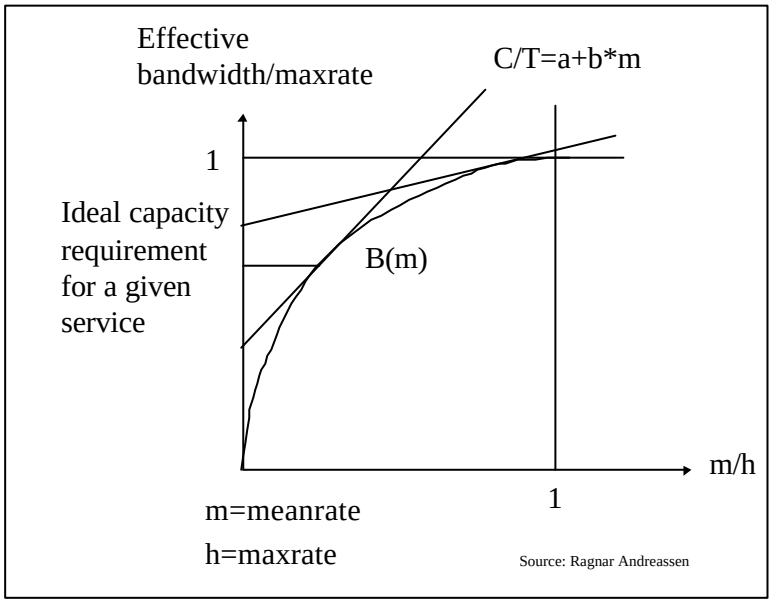

Figure 9. Kelly's abc charging

Thus, improved predictions of the traffic characteristics of an application leads to decreased costs per unit time, since the effective bandwidth is reduced. The concavity of the functions representing effective bandwidths is related to the peak rate of the sources. The higher the peak rates, the more concave is the function. That is, the incentive to accurately estimate the

[27] There will probably be some kind of policing function involved.

mean rate of a call increases. If the peak rate is low it will simplify statistical multiplexing and hence the need to predict the mean rate exactly becomes less important.

9.1.4 Charging for Content

In Sweden, the supply of content is not very extensive compared to countries like Japan. There, i-mode provides hundreds of accessible applications offered by equally many content providers. Along with the evolution towards more resource requiring applications, the question appears whether these are to be charged based on capacity consumption or perceived user value. This may be crucial for the success or failure of the UMTS business case.

A service provider could buy content and resell it to the subscribers or arrange access to it. The revenue is shared between them, or else the service provider could simply add a margin to the content price. It seems as if application-based charging is more flexible compared to pure usage base charging, which is the same irrespective of content. The reason is that applications consume radically different amounts of network resources. That does not mean that the applications are very diverse when it comes to perceived consumer value. SMS is a very good example of a service, which consumes negligible amounts of network resources. Still, the price for a maximum of 160 characters is in the order of 1 SEK. If applications like music or videoconferences were to be charged accordingly, they would never enter the market for content.

9.2 M3I (Market Managed Multi-Service Internet)

The M3I is an ongoing project under the European Union's Fifth Framework Program[28]. Its objective is to investigate a system for Internet Resource Management. Specifically it enables differential charging for multiple levels of services, offering the customer several choices of price, quality and reduced congestion.

The markets for differentiated services could be more effective since tariffs are communicated and validated in real-time. An application requiring a high level of QoS could be charged accordingly and bad congestion effects are avoided. The model illustrates that any user is allowed to send large amounts of bursty traffic, without compromising the right of any other user to do the same thing, as long as the user pays a premium for that opportunity. The need for priority queues, large buffers within the network and acceptance control at the border is eliminated. The end nodes convey the congestion status in the network (www.m3i.org)

The M3I project presents some direct benefits for both customers and network providers. Providers will face stronger competition, which means better price and QoS offers. There will also be a real-time feedback of the charges. The provider will be able to charge the user dynamically to encourage flexible usage of existing network resources.

Overall, the project illustrates a three-tiered model, including an Enterprise Policy Layer (business cases), an Application and Middleware Layer (offer, maintenance and update) and a Service Provisioning Layer (including the technical infrastructure) (M3I 2000b).

[28] Participants: Hewlett-Packard Ltd, (Coordinator), BT Research, ETHZ, Darmstadt University of Technology, Telenor and Athens University of Economics and Business

9.3 The M3I Architecture

This section investigates the charging architecture proposed in the M3I project. First, some general principles are described, which are considered to be fundamental for the further presentation. The architecture is a template model, which tries to illustrate the philosophy of real-time charging and pricing (M3I 2000a).

- Granularity: When possible, the charging and the service granularity should be the same, i.e. if services are provided on a per packet basis, charging should be based on the number of packets consumed.
- Commercial openness: Protocols related to commercial decisions must be implemented at the application layer, in order to give the service providers the opportunity to differentiate themselves.
- Edge pricing: Each contract (pricing, responsibility of failure, charge advice etc.) is focused on a single provider-customer interface. Hence the customer is given an incentive to behave as the provider desires through price control.
- Scalability: Compute intensive operations should be distributed as far to the network edge as possible, in order to enable many different compositions of systems.

In the model, the traffic could go in either way between different actors, and the network service could be seen as a function that routs and forwards traffic.

The different network functions could be dedicated to a single customer, but it seems likely that many persons in the same market segment have the same preferences. In this case, the functions are identical to many users with similar requirements. Consequently, the provider's enterprise policy will presumably take many customers' demands into account when establishing a certain policy. The same is of course true for the customer's policy function, which take more than one provider's offer into consideration.

The business interactions between customers and providers are managed on the business level. It even handles tasks such as pricing and service plans. On the application layer, communication services are provided and charged for according to customer demand (M3I 2000b). This layer abstracts away from the technical details of the network, which is taken care of by the Service-provisioning layer.

9.3.1 Usage Cases

The edge-centric usage case starts with an offer (one per market) from the service provider. It also establishes the pricing of these services and how it may change as demand varies. The offers are placed in a public offer directory. In parallel, the customers set their own buying policy for the applications in question. This is done in the customers' policy agent.
The agent configures a price reaction function along with its QoS control policy for a particular service and tariff. On the provider's side the charging and accounting system is configured according to the offer acceptance.

When the user starts using the network, the QoS manager keeps the traffic within its QoS policy, thanks to a metering function. The measured load is reported to the price reaction subsystem, which in turn regulates the QoS policy. The provider also performs metering in order to know how much to charge the customer. The revenue obtained is reported to the

price setting function, where prices change dynamically depending on the network load (M3I 2000a).

There is also a so-called edge-control usage case, where the main difference is the lack of charging activity at the customer. The customer instead relies on a specific charge advice feedback, from the provider. Only the provider meters the traffic, which means that the customer has no chance checking its accuracy (M3I 2000a).

9.3.2 The Inter-Network Usage Case

The inter-network usage case illustrates how market control can be found between two network providers. Taking consideration to more than one network provider seems realistic, since it is probable that the number of network owning actors in the future will exceed one.

In this scenario the QoS is not altered instantly. Instead the price reaction function is forwarded to the next provider's price setting function. The next provider is treated as a customer in the eyes of the first provider. However, when the service is forwarded (now to the end customer), it takes the role of a provider. The set-up of enterprise policies is similar to the previously described cases. They will probably be phrased in terms of aggregate measures, compared to single packets or flows. It is also probable that differentiated prices will be published, depending on traffic class etc. (M3I 2000a).

9.3.3 The Risk Broker

The risk broker is an agent, who takes the full end-user relationship from the network provider and offers charging, pricing and service interfaces. The risk broker re-sells network services and monitors the traffic flow in order to affect the QoS. The risk broker looks for the best offer from a number of network providers. This could be realized by the operation of a router in order to switch providers depending on the most beneficial price.

The broker sets its own pricing policy and forwards an offer of guaranteed service. The sold capacity appears to be reserved, but it is in reality a combination of services, which the broker buys for congestion-based prices from service providers. The longer the customer is connected, the greater the risk for the broker. Thus prices per unit time increases. Alternatively, the risk broker can offer a constant price independent of the duration of the session that on average covers the risk.

The message sent from the customer, asking for a certain QoS, reminds of that for an RSVP message, but it is implemented as an application layer protocol. In summary, the risk broker adapts two different charging schemes. The one representing its provider aspects is based on the principles of the 'abc' tariff (see above). The other, which monitors its role as a customer, is represented by a congestion pricing mechanism. The risk broker's customer-side charging system is logically separated from the provider side (M3I 2000a).

9.3.4 The Clearinghouse

The clearinghouse usage case involves at least five stakeholders: the ASP (application service provider), the end-customer, two network providers and finally a clearinghouse.

The providers offer their prices to both the end-customer and to the clearinghouse, whose primary business is to accept offers from multiple providers and retail it as a combined end-to-end service offer.

The clearinghouse might take responsibility for the network service it retails. If the service degrades for some reason, it intermediates between the provider and the customer. Perhaps the customer is given refunds, while the provider is requested to improve the network performance.

When the ASP has recognized the offer, it releases its own offer to the end customer who is charged both for the QoS transmission and the application usage. The latter charge may be based on the time spent playing a network game or the logs of a video server for example (M3I 2000a).

9.4 CAS (Charging and Accounting System)

This section will give a brief description of the charging and accounting system proposed in the M3I project (M3I 2000b). In order to make this a reality, lots of technical components are needed. However, the CAS architecture must be service independent, to enable future extensions of services.

Three different levels of charging are included in the model. First, there is a transport charge (access charge), which commonly is based on the underlying infrastructure. Second, the service charge is located on top of the transport charge and allows a distinction of separate service classes, with different QoS requirements. Third, the content charge relates to the accounting of information, such as reading or copying, which needs to be paid for.

The scenario deals with interconnected service providers, whose networks include this equipment. The model is so far insufficient, since many issues including for example multicast and security problems are outside the scope of the project.

The ISP's network consists of metering, accounting, charge calculate and billing functions. The metering could be handled by single units in the network, but could also be integrated in the routers. The measured traffic, which most likely is mediated into a more suitable form, is sent to the accounting system. Multi-service accounting must be supported and therefore the incoming data must be separated according to service classes. Thereafter, the information is forwarded to the charge calculate function, where the charge is obtained, by looking up the price for the utilized service. M3I (2000b) lists further characteristics, among other things the ability to present predictability of charges, allow for different business models and fraud protection.

9.5 Conclusions

Clearly, different charging schemes could be used for different purposes and all of them suit some applications and services better than others. A general perception is however that the importance of usage based charging will dominate in UMTS networks.

The number of chargeable events will increase, along with new actors on the market for mobile communication. Moreover, customers' loyalty to single providers is predestined to decrease. Therefore, subscriptions might become less useful as charging bases.

Much focus is put on how to charge the customer in real-time. How this is actually tried out in detail is so far not clear, but there are some statements on how it will be done. If a scheme is based on the mean rate of a traffic flow, there will be no incentives for the user to limit the max rate. Similarly, if the scheme is based on the max rate, there will be no incentive to limit the mean rate. Effective bandwidths do in fact take both these restrictions into account and determines the ideal capacity requirement for a given service.

The M3I project is one of the most rigorous experiments so far regarding Internet charging. It illustrates that charging must be carried out through mutual agreements between the customer and the provider. Thanks to inherent features of dynamic pricing, the demand for service will always equal the current supply thanks to the price mechanism.

10. Pricing

Enormous amounts of money have been spent on the 3G business case including licenses and the planning/development of terminals and systems. The market approach regarding GPRS is hence utterly crucial for the success or failure of the UMTS. It is vitally important to price usage in a way that corresponds well to customers' willingness to pay for the services.

GPRS applications and their corresponding charging are motivated starting points in several senses. First, an adequate course of action could be to compare what effects GPRS charging will have on traditional mobile telephony. Second, it reveals the prices of future applications, given today's charging structure.

10.1 GPRS Pricing

Europolitan recently released two new subscription types, i.e. GPRS WAP and GPRS surf[29]. GPRS WAP suits persons who primarily use their phones for small transmissions via a WAP server. If the volume of transmitted data is comparatively larger, GPRS surf is recommended. The actual transmission rate varies during the session, since resources are allocated dynamically, depending on the number of free GSM channels.

Prices related to GPRS usage (Europolitan), consist of a subscription fee, a fixed monthly fee and a fee for the data volume transmitted according to Table 11 (www.europolitan.se).

[29] These are primarily focusing on the consumer market, but GPRS could also be chosen as a complementary service to 'Business Partner', 'Företag 2.2' and 'Europolitan'.

	Start fee for conversation	Price conversation[30] SEK/minute	Price data SEK/MB	Monthly fee, SEK	Additional fee, >24h per hour, SEK
GPRS WAP					
workdays					
(07-19) o'clock	0.32	1.82-1.98	140	175	2
(19-01)	0.32	0.90			
(01-07)	0.32	0.55			
Other days					
(07-01) o'clock	0.32	0.90			
(01-07)	0.32	0.55			
GPRS SURF		See above	20	295	2

Table 11. Prices for GPRS usage

Charging and pricing still constitute unclear factors in the early adoption of GPRS though. For example, announcing the price for reading the newspaper turns out to be speculative. According to Europolitan, an hour of "wap-surfing" corresponds to about 0.1-0.2 megabyte of downloaded data, depending on the amount of information on the actual site (Lindqvist 2000). The price for such usage corresponds to 14-28 SEK in case of a GPRS WAP subscription and 2-4 SEK with GPRS surf.

In Sweden, people generally request useful applications, such as e-mail and other types of messages. An email sizing 4 KB (without unnecessary HTML code and attachments) would cost 0.53 SEK and 0.076 SEK, with WAP and surf respectively. At a first glance, these prices seem reasonable.

10.2 Pricing and Traffic Traits

Applications, including for instance streaming services, require strict support of quality of service (see above). It is not apparent how operators/service providers should price such actions. There are many reasons why this question ought to be further investigated (Ericsson & Persson 2000):

- The resources consumed do not correlate well with today's prices for packet-based services.
- There must be some match between perceived customer value and actual prices.
- The flow of money between the actors must be distinguished.
- The extension of QoS differentiation between market players is so far unknown. Hence, competition about customers, related to quality issues, will probably arise.

[30] Connections to operators other than Europolitan will suffer from an additional switching cost of 2.90 SEK.

- Some applications demand specific service classes. Hence, there must be a supply of alternative service classes and different prices for those. The effective-bandwidth charging scheme serves as good base, since the customers are encouraged to choose the optimal service class. (See above).

- The trend towards bundling (i.e. you pay one fee for a portfolio of services/applications) must be closely analyzed. The bundling of applications, which radically differ regarding resource consumption into one service class, may in fact be ineffective to corporate revenues.

Applications and services, described in section 5.2 above, point at the problem of combining resource requirements, QoS issues and pricing. Table 12 presents the prices for using (streaming) the applications over today's GPRS networks (Ericsson & Persson 2000), (Engman 1999). Even though, the example stands out as unrealistic, due to the comparison of existing prices and *future* applications, it serves as valuable information open for further analysis. The results are based on the GPRS surf subscription (fixed fees excluded). Obviously, these applications will never enter the market, given the prevailing prices.

Applications	SEK/Minute
Real time audio streams (CD quality)	1693
Mp3 sound (CD-quality)	154
Mp3 sound (radio-quality)	67
Mp3 sound (telephone-quality)	10
MPEG-1 video coding	2232

Table 12. The price per minute for a couple of streamed real-time applications

One could argue that the price per bit has to decrease in the future in order to make certain usage attractive. However, this may not be the case, since the price per bit is predicted to stay constant, even in a UMTS context. If the bandwidth is increased hundred times, the corresponding cost will increase hundred times as well (Zirn 2000).

Assume that music files are downloaded and stored at the terminal instead. This alternative does not seem too appealing either, given the prevailing prices. An average mp3 file is about 5 MB, which corresponds to a price of 100 SEK/track. Say, the price per MB would be 3 SEK. In this case, a song would cost about 15 SEK to download, a rather "fair" price compared to that of an ordinary CD record.

This discussion highlights the importance of pricing services according to perceived utility. Next section deals with the notion of cost vs. value based pricing.

10.3 Cost Based vs. Value Based Pricing

Content in general is priced in a way that does not equal the cost of providing it, due to peoples' willingness to pay for it. The example often referred to is data transmitted over SMS, which accounts for about 7% of the operators' incomes (Affärsvärlden 2001). Evidently,

customers pay for some indefinable perceived added value (see section 6.1.1), which turns out to be appropriate as long as the price for the service covers the cost of providing it.

Certain services of future interest are so far not compatible with the UMTS environment. This requires a thorough study of feasible ways to determine the correct price. Such a discussion is performed in Ericsson & Persson (2000), considering fixed Internet solutions. The logic of it happens to be equally applicable to this work. In Table 13, some of those pricing examples are listed with associated characteristics, benefits and drawbacks.

Price Method	Characteristics	Benefits	Drawbacks
Markup price	$\frac{unit\ cost}{(1 - ret.o.sale)}$ Unit cost = variable cost + fixed cost/units sold	• All customers face the same price. • The cost is easy to determine.	Since the current demand, perceived value and competition are left without consideration, this method will hardly lead to an optimal price.
Target-return price	$\frac{unit\ cost * ROI}{unitsales}$ ROI = Return on investment	The same as above	The same as above.

Table 13. Two pricing philosophies and their corresponding benefits and drawbacks

There are some general risks associated with cost-based pricing in general. First, by adopting this pricing method there is a chance to lose ground to competitors having lower costs.

Ericsson & Persson (2000) (p.28) also point out that "…if one services class generates a higher added customer value than the difference in cost for providing the service, then strictly implementing cost-based pricing causes the company to forego an opportunity to obtain higher revenues."

The notion of value-based prices will probably receive more attention as pricing is cleared in a real-time manner. With that, prices could be adapted to the current customer demand. The fact that services are getting more personalized also support this statement. The operators can find out about customer's willingness to pay for a service thanks to price discrimination. How this is performed is further investigated in 10.7.

Next section will look at budget constraints and people's willingness to spend money on mobile services.

10.4 Budgets

Individual preferences (see section 5) and budget constraints determine the actual behavior of the customer. The outcome coincides with the highest feasible level of personal satisfaction. According to chapter 3, there are numerous actors on the market for wireless communication. All of them exist for one reason only: to optimize revenues. In each step of the value chain a charge is added, that is, the price for usage in the eyes of the consumer consists of several "sub-prices" added along the way. Figure 10 illustrates the margins added by content, service and the network providers.

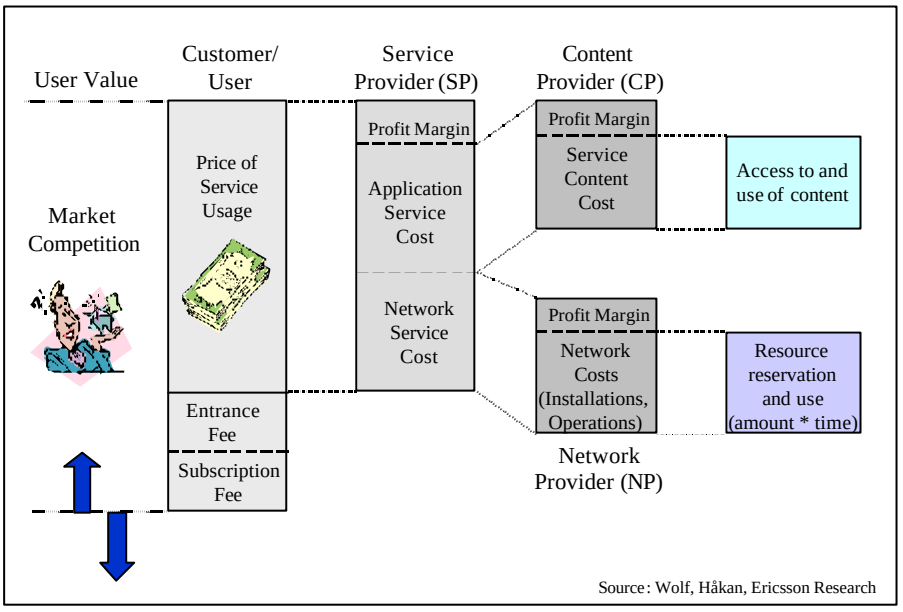

Figure 10. Each actor's share of the customer's spending

As can be seen, customers possess a certain budget, which is aimed for the expense of communication services. In the short run, this can be treated as fixed. Though, surveys, such as Mörk & Wennerström (2001) support the hypothesis that the average monthly bill increases, due to the introduction of data services. Factors such as the overall level of national inflation and the percentage change in public incomes could perhaps affect spending in the long run.

In order to price future UMTS services correctly it, could be instructive to learn from customer demand and price levels of GSM telephony and GPRS services. According to (Europolitan (1999/2000), the average revenue per user and month, in the beginning of year 2000 was somewhere between 500-600 SEK and month. For NetCom (1999), the average income per customer 1999 was determined to 418 SEK. The cash flow comes from different types of services (section 5) and the revenue streams dependent on customer profiles (that is business and private customers).

Still, there is no experience or feedback from the demand for GPRS services. According to a market research performed by Europolitan, customers in general think that an appropriate spending on GPRS services lies in the range of 150-200 SEK (Lindqvist 2000). However, if the price listed in section 10.2 is studied once again, this budget seems to be insufficient.

Now when we have looked at customers' propensity to spend money on communication services, something completely different will be presented, namely the competition between different market actors and its effect on prices.

10.5 Competition Between UMTS Operators

Pricing is closely connected to the market environment and the prevailing number of actors. Telecom operators have traditionally been characterized by monopoly, something that undergoes changes.

Ericsson & Persson (2000) describes several factors that favor large operators in their role as Internet access providers, at the expense of small ones. These are some advantages:

- When price and value for access decrease due to competition, the importance of providing transit increases.
- Competition leads to marginal cost pricing. Large operators have low marginal costs, a favor in an industry that is characterized by high fixed costs.
- It is easier to provide for QoS on your own network, instead of being dependent on interconnections.

The number of Swedish future mobile operators will be quite small, due to the limited amounts of UMTS licenses[31]. This means that the market for mobile Internet accesses will hardly be characterized by perfect competition (see Appendix A). As it is, the market is more like an oligopoly, with a streak of strategic behavior. The prices adopted by one actor immediately affect those taken up by others (see Appendix A). This is achieved by so called 'tacit agreements', where operators come to a common agreement, without actually discussing it among them. This is called self-enforcing agreements, i.e. each firm finds that abiding by the agreement is in their self-interest (Katz & Rosen 1998).

The market is in a so-called Nash equilibrium when each operator chooses the strategy that maximizes its profit, given the strategies of other operators in the market. Nash equilibrium among firms that choose prices is known as Bertrand equilibrium. This could be explained with a simplified example, using two operators A and B. Assume that A offers its services for a price P_A, then B maximizes it profit by setting the price to P_B, and vice versa. In order to determine the Bertrand equilibrium, a good starting-point is the market demand curve. If the market is characterized by oligopoly, the demand for one operator's services depends on the other's behavior. Three main states are identified:

1. Operator A's prices are higher
2. Operator B's prices are higher
3. Both operators' prices are the same.

In the first case all subscribers will choose operator B, which will satisfy the entire market demand. The same is true if operator A is the cheapest. In case both subscribers' prices are the same, they will satisfy half of the market's demand each (given equivalent services).

[31] MVOs are not considered.

Prices must of course be higher than the operators' costs. The question is then, if there is a price higher than the marginal cost (MC), that constitutes a Nash equilibrium (See appendix A). The answer is yes, in case both operators' prices are the same. This situation is not sustainable, since the objectives' of A and B respectively, is to maximize profits. Some of them would try to undercut the other's price, which would finally result in marginal cost pricing.

10.5.1 Scale Effects

The subscribers in a UMTS network can enjoy bit rates of up to 2 Mbps (shared medium). When the number of users increases within an area each of them suffers from service degrades. Therefore, the user (and the operators) may earn benefits from sharing agreements thanks to scale effects. Foros et al. (2000) illustrates a duopoly case, i.e. a strategic behavior between two actors. Two operators determine the infrastructure investment and then compete with each other under Cournot conditions (see Appendix A). The focus is on what effects a roaming policy has, both when regulators decide on roaming quality and when the operators themselves do it. The model is built around game theory.

A justification to the assumption of Cournot competition is the scarce radio spectrum and the fact that there are both technological and physical limits to capacity. The quantity in question is the number of subscriptions.

According to Foros et al. (2000), some important conclusions are made regarding consequences of roaming agreements. For consumers, an increase in roaming quality has two effects. Under competition, the quality increase leads to reduced competition and decreased market size. Given a certain quantity, prices for services will fall and so will the total quantity output. Hence, customers' welfare is affected both positively and negatively. By setting a high roaming quality, operators can restrict the amounts of infrastructure investments and with that the output in stage 3 in Figure 11. The equilibrium price increases as a result of this.

If a virtual operator is taken into account, which relies on an operator in possession of network infrastructure, the situation becomes ambiguous. The network-owning operator's incentive to invest in infrastructure may increase due to spillovers. This case differs from that without virtual operators. The reason is that investments stimulate innovations, and the customers are offered improved services and applications (Foros et al. 2000).

10.5.2 Competition Between Content Providers

The future number of content providers is in contrast to UMTS operators very difficult to estimate. Content can take many forms and the actors do not directly compete with each other. This is because the market is characterized by continuous flows of new players, all with different purposes[32]. The type of competition can hence be signified by the notion of monopolistic competition, i.e. that each actor can act as a price maker. This could be motivated by that most services are imperfect substitutes (see appendix A).

[32] See i-mode.

Thanks to 3G networks, the variety of available content is predestined to develop dramatically. However, the evolution towards the UMTS may take some time and to begin with, the number of actors may be characterized be a fixed number, leading to a short-run equilibrium (see Appendix A). The economic profit reminds of that of the monopolist, but in a smaller scale.

The long-run equilibrium differs somewhat from the short-run equivalent. New providers will enter the market as long as there are profits to be made. The question is how new entrants will affect the market equilibrium. The firm specific demand curve will be shifted inwards. In the long run the content providers must earn enough money, so that the average revenue (price) is equal to the average cost. A representative provider must therefore "produce" at an output level where the demand curve and the average cost curve are tangent (see Appendix A).

However, in the end, there must be some way to find out which customers are prepared to pay for a certain service. Recognizing the behavior of different demographic groups could do this. Section 10.6 will look closer into this type of reasoning and specifically how to price discriminate between different user groups.

10.6 Price-Discrimination

When different people are charged different prices for the same service, they are exposed to price discrimination. This could be useful for the operators/services providers since consumer's willingness to pay varies. Thanks to price discrimination, the firm's profit will increase. However there are some conditions, which must be satisfied (Katz & Rosen 1998).

- The firm must be a price maker.
- The firm must be able to identify which customers is which.
- Consumers must not be able to engage in arbitrage[33].

Price discrimination is divided into three types, namely first, second and third degree. The first degree means that each unit could be sold at a price, which just equals the buyer's maximal willingness to pay for it. There are certain problems related to this behavior, since determining peoples' marginal propensity to consume turns out to be rather tricky.

In reality, it is impossible and therefore the operator could observe the customer's own actions as a basis for price discrimination (second degree price discrimination). The same prices are offered to all consumers, but they sort themselves out due to self-selection. Here, so called two-part tariffs serves as a good example. For example, there may be some fixed subscription fee, but there is also an additional per-use/per-unit fee. Given people's heterogeneous willingness to pay, an optimum combination of fixed and varying prices could be found. The idea is to keep as many types of users as possible.

[33] Arbitrage means that the customer whom the operators charge low prices could resell the product (service) to customers who would have to pay higher prices.

Finally, there is the third degree price discrimination, which in practice means that the operator could divide the overall market into market segments. The operator produces at a point where the marginal revenue equals the marginal cost in each market. Each market segment is approached with different types of subscriptions. For example, Europolitan offers business customers several alternatives whereof these are some of them (www.europolitan.se):

Type of subscription	Characteristics
Eurobusiness	• Companies with 50 employees or more. • The firm's equipment is taken into consideration.
Business Partner	• Services and prices are adapted to the company's habits regarding mobile phone usage. • The more usage the lower the price for each call.
Företagsavtal 2.2	• A company with a total call time of 100 min/day saves money. • The mobile phone could be used even as a fixed phone in-house.

Table 14. Different types of subscriptions offered by Europolitan

Hence, companies choose the subscription type that best suits them. Price discrimination proves to be a commonly forthcoming phenomenon, used in many different industries, not only in telecommunication circumstances.

10.7 Conclusions

The UMTS business case is partly dependent on the way the services in question are priced. Here, GPRS prices serve as the best information attainable at the moment. As can be noticed, prices for voice do not differ much from the prevailing prices for voice in ordinary GSM networks.

The appropriateness of the prices for data transmissions in GPRS networks is questionable. As far as the type of information is modest, the cost for usage is motivated. For example, an hour of intensive WAP usage costs about 14-28 SEK with a GPRS SURF subscription. Further, an e-mail of size 4 KB costs at most ≈ 0.50 SEK (GPRS WAP subscription).

The serious problem appears when dealing with resource demanding applications/services such as music downloading and the transfer of moving pictures. Given the prices for GPRS services, these types of activities will hardly be requested unless prices fall. Unfortunately, the price per bit will probably remain constant, due to infrastructure investments. That means that even if higher transmission speeds are admitted, thanks to UMTS, the services will remain very expensive. The perceived utility simply does not compensate for the high cost of usage. Here, alternative ways to charge the customer have to be considered. Hence, seen from a business perspective, the price per bit ought to fall in the future.

Sometimes, none of these charging methods are satisfying. This is due to comparatively large amounts of resource consumption that the customer is not prepared to pay for. The per-

ceived utility simply does not compensate for the high cost of usage (e.g. downloaded music). Here, alternative ways to charge the customer have to be considered. Of course, the same is true the other way around. That is, some services do not consume very much resource at all, but are still perceived as highly valuable to the customer. Hence, these services could be charged in a way that does not correspond to the cost of providing it (e.g. information services). Hence, the bits belonging to different applications/services are priced differently.

The competitive climate will also affect the prices. Since the number of network operators is small, an appropriate approach could be to characterize the market with an oligopoly. The immediate consequence is that the price charged by one operator would be the same for all operators (almost) and the price will eventually narrow the firm's MC. One possible way to differentiate the service is to offer a specific (unique) level of QoS.

Still, the number of content providers is large and the applications provided are quite dissimilar. Hence, they could perhaps be priced individually (high or low). However, there are few loyalty agreements with the customers and therefore the price may be forced down due to poor demand conditions.

Different customers could sometimes be priced differently. Often this kind of action is not explicitly expressed, but typical examples are different types of subscriptions and business packages, which target specific user groups. This is often called price discrimination since the same service could be priced differently, depending on who is using it. The concept of price discrimination will probably be further utilized in the future. This is clearly understood when studying the M3I project, where prices are adjusted in real-time, based on the customer's individual preferences.

11. The UMTS Business Case

Pricing and charging must not be studied separately, but in a broader context including the complete UMTS business case. The objective of this section is to investigate some trends and circumstances related to the deployment of 3G networks.

The aim with the section is not to go too deep into details. Instead it will try to explain some market traits, which complicates the procedure of pricing and charging mobile services. This part will serve as a complement to what has been written in the previous sections.

11.1 Strategy

The number of UMTS operators in Sweden will be limited. Today there are four competing licensees, which are permitted to run 3G networks. They are restricted due to the shortage of radio spectrum, which is a natural impediment to the establishment of more operators.

Alliances like the one recently announced between Telia and Tele2 could however be more frequent or even necessary in order to stay competitive on the future market scene. Immedi-

ate benefits from cooperation are expanded customer bases[34], which seems to be vitally important for new entrants and shared costs for the infrastructure. A common system could save as much as 50 % the cost for the actors' operation and maintenance, according to Mitch Lewis, general manager for multi-service networks at Ericsson (Newing, 2001a). However, operators such as Orange and HI3G[35] share a possible benefit in that they do not have to care about already existing GSM subscribers (Affärsvärlden 2001).

It is uncertain how quickly the revenue will take off, but one thing is completely clear: the operators must fill their networks with traffic as soon as possible in order to recover from costs associated with the network investments. Though, filling the network with voice is not profitable, since the price for voice is falling. Returns will come from high value data services, which in large extent will be provided by content providers like AOL and Yahoo. The mobile operators hope to take about 20 % of the services in the UMTS network and about 5% of the revenues from mobile commerce (Affärsvärlden 2001).

Experiences from Japan tell that the email will be the killer application, while entertainment will be the killer content (Cane 2001). According to the same source, more than 50 % of all user activity deals with downloading of network games, screen savers, phone ringing tones etc. It is uncertain though, if the western operators can match NTT DoCoMo's charging structure. Cultural differences and high initial costs constitute obstacles, which could prevent a successful replication of i-mode in other countries.

However, costs for infrastructure investments are associated with sunk costs, i.e. once made, they do not affect future cash flows. Pay back times and the return on investment must nevertheless be taken into consideration. Thus, this cost will in one way or the other be transferred to the customers. In Shillingford (2001), the pay back time for operators, who have obtained the license for free, is estimated to be around 7 years. This is one reason why operators will have to expand gradually.

The operators in Sweden are obligated to cover 99 per cent of the population (Shillingford 2001). Base stations are only one part of the total cost. Revenues must also cover costs like network set-ups, acquisition of content, development of products and applications, customer acquisitions, handset subsidization and marketing costs. Unfortunately, small operators may find it hard to enter the capital and equity markets. Perhaps, the only available alternative would be to merge with larger, more credit-worthy rivals.

Guesses about the market structure, size and number of core actors are very risky to do. There are so many parameters that tend to influence different outcomes. Often, there are other components, which control the business landscape, than just the relationship between different actors and their strategies.

An immediate threat against the success of 3G is Wireless Local Area Networks (WLANs). Thanks to the decreased cost of this technology, it has become more commonly used within firms and by residential users. The major drawback considering WLANs is the short reach of approximately 100 m, but the admitted transmission speed is much higher and there are no

[34] Telia and Netcom have about five million customers together (Affärsvärlden 2001).
[35] Owned by Hutchison and Investor.

restrictions regarding spectrum utilization. Thus, the usage of these networks is cheaper (Affärsvärlden 2001). However, a major drawback associated with WLANs is the poor ability to guarantee QoS, due to the lack of licenses.

11.2 Factor Conditions

As mentioned above, the construction of UMTS networks will impose lots of challenges on every single operator. Until now, the most debated question has been how to finance the needed infrastructure. The supply of it is still taken for granted, which gives rise to uncertainties. The number of demanded base stations will be four times of those in the current GSM system, according to Ogelid (2000). Consequently a great part of the UMTS services will not be available on the market in the short run. Even if the starting dates are achieved, experts mean that services will be limited until at least the middle of 2002 (Mcartney 2001).
Customers must, to be able to use UMTS services invest in new terminals. Seen from an overall business perspective, this ought to profitable. There are however severe requirements on the terminal developers to work on the same time scale as the suppliers of infrastructure.

Further reasons to delay in UMTS services could be the lack of well-established industry standards. The concept of WCDMA may be immature and it is very likely that manufacturers will interpret the features in different ways. In the worst case, the terminals of one supplier will not be compatible with the network of another.

11.3 Demand Conditions

In order to acquire knowledge about future demand conditions, operators ought to investigate which types of applications that seem most interesting. As can be noticed in section 5, most real-time applications will not be demanded until the infrastructure can handle the corresponding data rates.

UMTS will simply add more capacity, and past experiences say that capacity, no matter how much, will be consumed in one way or another. Suddenly, you may find the application that strictly speaking kills everything. Hence UMTS are better dimensioned for such a scenario than any other technology.

However, there are still major risks associated with the network (network business), mostly caused by inaccurate demand forecasts. Some of the key uncertainties, brought up in Elvidge & Peirce (1999) are derived to:

- Empty networks and platforms, generating little or no revenue, representing fixed costs, which are not paid for.
- Networks that are full at busy time, but idle for most of the day, in which case the revenue from the busy period may or may not cover the costs.
- Adverse reaction from customer business units, following jumps in unit costs as a result of unforeseen but necessary network enhancements.
- Low actual customer demand compared with forecasts, hitting the profitability of an individual service.

- Inflexible networks, which are unable to support new service proposals.
- Networks that have no headroom to accommodate new services, or growth in existing services.
- Reduced profitability of service affected by jumps in unit costs imposed by the network business.

Despite the above listed issues, there is need to figure out how long it will take for people to recognize the inherent benefits. As previously stated, the infrastructure requires lots of provided capital. Investors and operators are forced to cut the payback times as much as possible, covering rates, amortization and an adequate return on investment. It should also be announced that terminal developers as it is are not eager to produce new technologies if there is too much risk involved, as they are making so much money out of existing equipment.

Customers must be given incentives in order to start using services over UMTS. How this is realized is not surprisingly arguable, since EDGE could in fact offer well enough capabilities in order to satisfy customer needs. It should be recognized though that the QoS enabled by UMTS is better than the corresponding level achieved by EDGE.

This concludes the fact that operators must be eager to encourage content providers to offer customers services and applications. Perhaps sharing incomes from users in a greater extent than is the case today could do this. It is dubious however whether telecom operators will tolerate any loss of traffic income, since they are meant to finance the traffic growth and future mobile network structures.

Ericsson believes that a move from voice to data-centric wireless services will generate strong demand for new mobile infrastructure (Parkes 2001). Voice is the most wanted mobile service so far and to the extent this will change within the nearest years is yet to be discovered.

The next section will describe related industries and their support of the UMTS business case.

11.4 Related and Supporting Industries

Most concerns regarding UMTS are dedicated to the difficulties in finding wealthy investors, willing to share the expenses of building the network infrastructure. To begin with, UMTS operators will draw from the already existing GSM infrastructure, asking for additional antennas in order to make the system both 2G and 3G compatible (see section 4). This solution is only temporary and new equipment will eventually be required. Then the question arises whether the facility-based operators should lead the construction work alone, together in joint ventures, or outsource that business to a third party.

The first alternative will undoubtedly result in huge up-front costs and the dependency on high prices for usage will increase. There are obvious drawbacks related to this solution regarding the business profitability, inflated payback periods, and concerns finding prospective sources of finance. Moreover, the actions of competitors tend to be a question of vital importance.

The second alternative seems to be more attractive. An example of current interest is the proclaimed cooperation between Telia and Tele2, which admits certain benefits to both actors. The combinations of a huge customer base, shared costs and the control of UMTS frequencies seem very attractive. The rise of MVOs (described in section 3) could probably lead to valuable partnerships, both regarding the way costs are split and the way customer demand is enhanced.

The last alternative will probably not be forthcoming in Sweden, but the idea is to delegate or outsource the construction of towers etc. to another actor. These so called site owners could focus on developing competitive advantage and cost-effectiveness in the business and quality aspects of the physical equipment (Newing 2001b). The operators should instead concentrate on its intellectual property in radio optimization.

It is not apparent how operators are supposed to manage the finance of 3G networks. Vastly increased liabilities make the operator vulnerable to failed customer incomes. Besides, it is not even sure that the money will be there for them. However, the problem could be solved thanks to vendor finance. That means that the supplier (e.g. Ericsson) is paid only part of the value of the contract. The rest is treated as a loan to the operator, secured on the value of the equipment. The loans are subsequently transferred to banks or the capital market. This process however puts serious strain on the vendor's balance sheets.

11.5 Conclusions

Alliances between operators seem to be a good way to save costs and earn benefits in UMTS networks. The major reasons for that could be:

- Costs are cut (possibly by 50%) concerning operation and maintenance.
- The customer base is increased.

Perhaps the demands and requests from already existing GSM customers could be a drawback.

The payback time for network investments of about 7 years demands a strategy, which aims at regaining as much of the infrastructure expenses as fast as possible. This is achieved by filling the networks with content, which generate high returns. Filling the network with voice is not profitable, since the price for voice is falling[36].

Another question of major importance is how the financing of the networks is supposed to be arranged. There are still uncertainties whether operators will be allowed to borrow the money required in order to manage the work of construction. Perhaps the existence of alliances will benefit the operators in some sense. Other alternatives, which have been considered, is vendor financing or outsourcing.

[36] However, the number of subscribers will rise. This could eventually compensate for the fall in price.

Finally, the move from voice to data centric wireless services is believed to be one reason why the demand for network capacities has increased.

12. Conclusions

Traditionally the network operator has been the receiver of most part of the revenues from mobile communication. People's demands for content will increase and the facility-based operator will not possess all skills needed in order to fulfill their satisfaction. At the same time, the costs associated with the UMTS infrastructure must be financed in some way. Both maximizing the network utilization and the revenues do this. Hence the role of content providers will probably become more important since the margin on content is more profitable compared to ordinary voice. Experiences from i-mode tell about content providers that generate a lot of value but no revenues. This must change in order to make the UMTS business case successful and profitable.

When it comes to what services/applications that are preferable in the context of 3G networks, there is balance between customer value and resource consumption. Some services (e.g. streaming services or resources requiring an application available through downloads) are believed to meet severe resistance in UMTS, due to the high price per bit. There will also arise complications regarding how to guarantee the user a certain level of QoS, with respect to delays, jitter and losses, especially end-to-end. This requires well-established agreements between several networks regarding the way traffic should be treated and how to map between different QoS enabling technologies. Thus, the number of requested video sessions are assumed to be quite modest, but the volume generated could be comparatively large. Hence, in order to make the network business profitable, the price per video session would have to be higher than most people are prepared to pay.

Services/applications, which include significant perceived user value, without heavy resource consumption (e.g. emails), will probably become more attractive seen from an operator's perspective. Here, algorithms, which aim at allocating the appropriate amounts of network resources, could be valuable when optimizing the network utilization. The choice of network technology also seems to matter. It is much about balancing the inherent stability regarding QoS guarantees in circuit switched networks against the improved flexibility in packet switched networks.

However, predicting people's demands for services seem to be a very hard task. Factors such as age, profession, IT literacy and the personal budget seem to play dominating roles. By carefully analyzing the market and the existence of complementary and substitute services, demand curves could be sketched and with that the ability to estimate price-elasticity is enhanced. However, by adapting dynamic pricing and charging in real time there is always an equilibrium point revealing people's willingness to pay for a given QoS. How much this is in reality is very much a question of the personal budget. Today, the average spending on mobile communication is somewhere between 400-600 SEK/month. However, experiences from i-mode illustrate that the average spending increased thanks to the introduction of data services.

The type of competition between operators and content providers further determines the actual prices for usage. The few number of licenses favors the assumption of a market characterized by oligopoly. This in turn means that prices will narrow the marginal cost of the network service. Actors such as mobile virtual operators could affect the competitive climate within the branch of mobile communication. The main benefits for the operators (and the customers) could be shared costs, spin off effects and enlarged customer bases. Possible threats could be increased competition and lowered marginal.

Content on the other hand could be priced in a way that vastly differs from the marginal cost of providing it. The individual propensity to spend money on content could be found out by different types of price-discrimination.

Finally, prices often reflect operators' needs to generate positive cash flows and a means to finance the construction of networks. The predicted pay back times for these investments will determine the price for services and vice versa. Consequently, the choice, whether, the UMTS infrastructure should be financed by credit institutes (e.g. banks), vendors (e.g. Ericsson), outsourced (probably not in Sweden) or a combination of them will almost surely affect the final price.

13. Further Work

There are several areas in this report that are not or merely briefly discussed. A deeper study of these seems motivated in order to understand the economy of wireless networks better. Here are a few examples that deserve more attention in future works.

- Today it is unsure what tomorrow's networks will look like. It could therefore be interesting to study the business opportunities associated with different network structures.

- The relation between the price and demand for a single service is often diffuse. What is needed is a framework that is able to handle the estimation of demand curves in a consistent and structured way.

- QoS is assumed to be a crucial factor considering how network operators differentiate themselves towards competitors. How this could be arranged and what effects this will have on charging and pricing must be investigated in detail.

- Despite many fascinating proposals of charging algorithms, very few of them could be implemented. Often they are too complicated seen both from a technical and a customer point of view. What is requested is a functional, easily understood algorithm that takes as many interests as possible into consideration.

- Today, the prices for services/applications differ considerably in fixed and mobile networks. The interpretation of the convergence between them and what will happen to charging and pricing is still an open question.

- In what extent is UMTS possible to combine with Wireless LANs?

- The demand curve for a business user does not always correlate with the one for residential users. A better understanding of inequalities between different demographic groups is needed and how to price-discriminate between them.

14. References

Altmann, Rupp & Varaiya. (1999a). *Internet User Reactions to Usage-Based Pricing,* Department of Electrical Engineering & Computer Sciences, University of California at Berkeley.

Altmann, J. Rupp, B. & Varaiya, P. (1999b). *The Case for Quality of Service on Demand - Empirical Evidence from the INDEX Project,* ISQE'99, Workshop on Internet Service Quality Economics, Cambridge, Massachusetts, USA.

Analysis, Affärsvärlden No. 3 January 17, 2001. *3G-Operatörerna-Enade vi stå…*

Andersson, J. & Wirde, J. (2000). *3G-324M Video Client/server,* Master Thesis at Ericsson (ERA/FZ/XF).

Barnett, S.A. (1998). *Billing for Services in Future Broadband Networks,* Telecommunications and Information Technology Research Institute, University of Wollongong Australia. IEEE.

Botvich, D. Chen, Y. Curran, T. Kerswell, B. McGibney & J. Morris, D. (1997). *On Charging for Internet Services provided over an ATM network,* IEEE.

Brandt, H. Eriksson, A. Evaldsson, P. Leijonhufvud, G. Ohlman, B. Olsson, G. Söderberg, J. Wedlund, E. (1999). *The Switchlab Paper on Real-time IP Networking,* Ericsson Switchlab.

Cane, A. (2001). *Key question as Japan's DoCoMo seeks to export i-mode to the west,* Financial Times Survey, Financial Times January 17, 2001.

Crawford, D.W. (1995). *Pricing Network Usage: A Market for Bandwidth or Market for Communication?,* MIT Workshop on Internet Economics.

Edén. K. & Arvidsson, Å (Ericsson Utvecklings AB). (2000). *Pricing of Competing Telecommunication Services,* NTS (Fifteenth Nordic Teletraffic) Seminar, Lund, 2000.

Elvidge, A. & Pierce, K. (1999). *New Products for Old: Making a Profit,* British Telecom.

Engman, A. (2000). *High Performance Video Streaming on Low End Systems,* Master Thesis, Department of Microelectronics and Information Technology (former Teleinformatics), Royal Institute of technology (KTH).

Ericsson, J. & Persson, P. (2000). *Charging for Internet Access,* Master Thesis, Department of Industrial Economics and Management, Royal Institute of technology (KTH).

Ericsson, J. Persson, P & Franzén H. (2000). *Services and Tariffs in Third Generation Cellular Networks,* Department of Industrial Economics and Management, Royal Institute of technology (KTH).

ETSI TS 122 115 *Digital cellular telecommunications system (Phase 2+) (GSM); Universal Mobile Telecommunications System (UMTS); Charging and Billing,* Draft V3.2.0, 2000.

Europolitan Holding AB. Årsredovisning 1999/2000.

Fenger, C. (2000). *The Self Similarity of Data Traffic,* NTS (Fifteenth Nordic Teletraffic) Seminar, Lund, 2000.

Fishburn P.C. & Odlyzko M.A. (1999). *Dynamic Behavior of Differential Pricing and Quality of Service Options for the Internet,* AT&T Labs-Research.

Foros, Ø. Hansen, B. & Sand, J-Y. (2000). *Demand-side Spillovers and Semi-Collusion in the Mobile Communications Market,* Scientific Report (R&D R 32/2000), Telenor FoU.

Foros, Ø. & Hansen, B. (2000). *Connecting customers and disconnecting competitors-The facility-based firms' strategy towards virtual operators.* Scientific Document (R&D N 73/2000, Telenor FoU.

Gibbens, R. J. (1996) *Traffic characterisation and effective bandwidths for broadband network traces,* Statistical Laboratory, University of Cambridge

Iversen, V.B. Glenstrup, A.J. Rasmussen, J. *Internet Dial-up Traffic Modelling,* NTS (Fifteenth Nordic Teletraffic) Seminar, Lund, 2000.

Jagau, A-W. (2000). *Developing Service Management In IP Networks,* ICM Conference, at the Garnd Hotel 21st & 22nd March 2000.

Katz, L.M. (University of California at Berkley) & Rosen H.S. (Princeton University) (1998). *Microeconomics,* 3rd edition, Irwin McGraw-Hill

Kelly, F. P. (1995). *Charging and Accounting for Bursty Connections,* McKnight, L. W. & Bailey, J. P. (eds.) *Internet Economics,* pp. 215-278. Cambridge, Massachusetts, USA.

Kelly, F.P. (1996), Malloo, A.K. & Tan, D.K.H. *Rate control for communication networks: Shadow prices, proportional fairness and stability,* Journal of the Operational Research Society 49 (1998) 237-252.

Kokkonen, V. (2000). *Managing Service Quality To Fulfill Customer Needs,* ICM Conference, at the Garnd Hotel 21st & 22nd March 2000.

Lindberger, K. (1999). *Balancing Quality of Service, Pricing and Utilisation in Multiservice Networks with Stream and Elastic Traffic,* ITC 16.

Lindqvist, C. (2000). *Få får tillgång till mobilt Internet,* Finanstidningen, December 1 2000.

MacKie-Manson, J.K. & Varian, H.R (1995). *Pricing Congestible Network Resources,* IEEE Journal on selected areas in communication, vol. 13, No. 7, 1995.

Marchent, B.G. Wilson, M. & Rouz, A. (1999). *Support of Mobile Multimedia over Radio for a Range of QoS and Traffic Profiles,* Fujitsu Europe Telcom R&D Centre, Stockley Park, Uxbridge, Middlesex, UB11 1AB, UK. IEEE.

Market-Managed Multi-service Internet (M3I). (2000a). *Deliverable 2 Architecture,* (The M3I Consortium: Hewlett-Packard Ltd, Bristol UK (Coordinator), Athens University of Economics and Business, GR, BT Research , Ipswich GB, Eidgenössische Technische Hochschule, Zürich CH, Technical University of Darmstadt DE, Telenor, Oslo NO), European Fifth Framework Project IST-1999-11429.

Market-Managed Multi-service Internet (M3I). (2000b). *Charging and accounting System (CAS) Design,* (The M3I Consortium: Hewlett-Packard Ltd, Bristol UK (Coordinator), Athens University of Economics and Business, GR, BT Research , Ipswich GB, Eidgenössische Technische Hochschule, Zürich CH, Technical University of Darmstadt DE, Telenor, Oslo NO), European Fifth Framework Project IST-1999-11429.

Mcartney, N. (2001). *Bottlenecks may delay roll-out programmes,* Financial Times Survey, Financial Times January 17 2001.

Morris, D. (Sherkin Technologies Ltd.) Pronk, V. (Phillips Research Laboratories). (1999). *Charging for ATM Services,* IEEE Communication Magazine May 1999.

Mörk, A. & Wennerström, P. (2001*). Prerequisites for Services over GPRS-Implications on traffic mix,* Master Thesis, Department of Industrial Economics and Management, Royal Institute of technology (KTH).

Netcom. Årsredovisning 1999.

Newing, R. (2001a). *All for one and one for all,* Financial Times Survey, Finacila Times January 17, 2001.

Newing, R. (2001b) *New service industry emerges,* Financial Times Survey, Financial Times January 17, 2001.

Odlyzko, A. M. (2000). *Content is not king,* AT&T Labs-Research.
http://www.research.att.com/~amo.

Ogelid, H. (2000). *Svensk teknik gör 3G-näten billigare,* Computer Sweden, December 4, 2000.

Ouchterlony, D. & Molin, M. (2001). *New Service Opportunities in 3G: A Relative system resource usage model for UMTS QoS classes,* Department of Teleinformatics, Royal Institute of Technology, (KTH).

Parkes, S. (2001). *Hugh Gamble' is paying off,* Financial Times Survey, Financial Times January 17, 2001.

Pehrson S. (2000). *WAP-The catalyst of the mobile Internet*. Ericsson Review No. 1

Philippopoulos, P.I. Georgopoulos, C.E. & Sykas, E.D. (1999). *QoS Interpretation in 3rd Generation Wirless/Mobile Systems,* Department of Electronical Engineering and Computer Science, National Technical University of Athens (NTUA). IEEE.

Prasad, N.R.. (1999). *GSM Evolution towards Third Generation UMTS/IMT2000, Lucent Technologies,* © 1999 IEEE.

Rouz, A. Wilson, M. & Marchent, B.G. (1999). *Broadband Interworking Architecture (BRAIN) for Future Mobile Multimedia Systems,* Fujitsu Europe Telcom R&D Centre Ltd, UK. IEEE.

Scott, A. & Irvine, C. (1999). *Mobile Data in Asia,* Merrill Lynch.

Shaoyan, W. & Chuanyou, L. (1998). *Research on the cost and tariff models of data communication service,* International Conference on Communication Technology (ICC).

Shillingford, J. (2001). *Site costs are set to soar,* Financial Times Survey, Finacial Times January 17, 2001.

Sung, N. & Cho S.H. (Management Research Laboratory, Korea Telecom). *(???). Telephone Fixed Rate Structure and Telecommunication Development,* Retrieved from http://userpage.fu-berlin.de/~jmuel…/sung_cho_telfixed_rate_structure.htm August 3, 2000.

Tamm, G. *Flitiga flickors tidsfördriv,* Kontakten (Tidningen för medarbetare inom Ericssonkoncernen), No. 20 (2000).

UMTS Forum. (2000). *The UMTS Third Generation Market-Structuring the Service Revenue Opportunities,* Report No. 9.

UMTS Forum. (1999). *The Future Mobile Market-Global trends and development with focus on Western Europe,* Report No. 8.

UMTS Forum. (1998). *The impact of licence cost levels on the UMTS business case,* Report No. 3.

Vrinis, T. (2000). Effective management of IP Quality of Service, ICM Conference, at the Garnd Hotel 21st & 22nd March 2000.

Wireless/Mobile Communication Europe. (2000). UMTS Part 2: *Forecasting Demand for Mobile Data Services in Europe,* Yankee Group.

Yang, J. & Kriaras, I. (2000). *Migration to all-IP based UMTS Networks,* GSM/UMTS R&D, Lucent Technologies, UK. 3G Mobile Communication Technologies, Conference Publication No. 471, © IEE 2000.

Zirn, T. (2000). *KTH professor: Genombrott för mobilt Internet hotat,* Computer Sweden, November 15, 2000.

Örnung, M. W. *Nödvändig bit i datapusslet*, Kontakten (Tidningen för medarbetare inom Ericssonkoncernen), No. 16 (2000).

www.citibank.com, retrieved on July 4 2001.
www.tullverket.se/valutakurser/21_31mars2001.htm.
www.nttdocomo.com/i/inumber.html, retrieved on March 25 2001.
www.gsmworl.com/technology/yes2gprs.html, retrieved on November 15th 2000.
www.europolitan.se (/upload/documents/PB.gprs.pdf), retrieved on December 15 2000.
www.m3i.org, retrieved on March 15 2001.
www.3GPP.org, retrieved on March 15 2001.
www.3gip.org, retrieved on March 15 2001.

Appendix A: Elementary Microeconomics[37]

The Demand Curve

The demand curve for a service shows the maximum quantity a user is willing to consume at any given price of the service, other things being the same[38].

The Cross Price Effect

The cross-price effect describes how the price of one service affects the demand of another service. The result depends on if the services are substitutes or complements.

Two services that satisfy similar needs are called substitutes. An increase in the price of one service leads to an increase in the demand for a substitute. If two services tend to be used together (i.e. complements), an increase in the price of one service means a decrease in the quantity demanded for the other.

One should differ between a *change in the demand* for a service and a *change in quantity demanded*. This relates to the difference between own-price and cross-price effects. If the price of a service changes, the quantity demanded will typically fall. The decrease is attained through the given demand curve. However, if the price of *another* service (complement or substitute) changes, the entire demand curve will shift (i.e. a change in the demand).

Marginal Revenue and Marginal Cost

The change in revenue due to the sale of one more unit of output is called marginal revenue (MR). The concept of marginal cost (MC) is of course analogous, that is MC is the change in the total cost due to the production of one more unit of output.

Consequently, the production of one more output will rise profits only if MR>MC. The marginal output rule says that a firm should produce at a level where the MR equals the MC. The Shut-Down Rule says that if for every choice of output level the firm's average revenue is less than its average economic cost, the firm should shut down.

Perfect Competition

Assumptions

Perfect competition is characterized by some fundamental assumptions. *First*, the sellers are price takers, since each supplier believes that its output choice has a negligible direct effect on the market price. Besides, each supplier thinks that it has no effect on the collective actions of other suppliers. If this condition was not satisfied, the action of one firm might trigger responses by other suppliers, which would collectively affect the market price. *Second*, the firms do not behave strategically; the firm does not anticipate any reaction by rival suppliers

[37] Appendix draws from facts found in (Katz & Rosen 1998). The information is however general.
[38] Other things are for example: the individual taste, income and prices for other services.

when it chooses its own action. *Third*, the entry into the market is free, which means that a supplier can enter the market without incurring any special cost. *Fourth*, buyers are price takers, meaning that sellers take prices as given.

Appropriate Market Structure

The market which best corresponds to the assumption of perfect competition is made up of many buyers, where each of them is small and unable to affect the market price. The number of suppliers is large and each actor has little direct effect on the total market supply. A large change in the output of one supplier will have little effect on the total percentage change in output. The service that the sellers offer is homogenous and consequently the price cannot be raised over the common level. That would only result in zero output. A market characterized by perfect competition also requires that buyers must be well informed about available alternatives. Figure 11 below illustrates the output and the corresponding price under perfect competition.

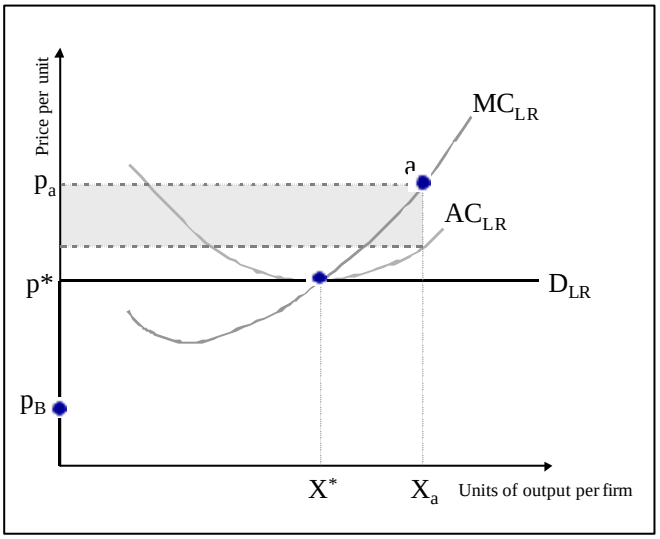

Figure 11. The output and the corresponding price under perfect competition

Monopoly

When the quantity that a firm buys or sells significantly affects the price the firm faces, the firm is said to be a price maker. The reason is that it can influence the price through its choice of output.

Assumptions

The demand curve for a price making firm slopes downwards, since the price falls as the amount of output rises and vice versa. Thus, firms do not have to behave strategically and the market entry is completely blocked. Moreover, buyers take the price as given.

Appropriate market Structure

A market characterized by monopoly does have a large number of buyers, where no one is able to affect the market price. The number of sellers on the other hand is only one, which

also means that there are no close substitutes. However, the customers are completely informed about the product or service.
The fact that the demand curve slopes downwards implies that the average revenue is falling, which in turn means that the marginal revenue pulls down the average revenue and lies below the demand curve. The firm still produces where the MR equals the MC, but unlike the price-taking firm, the monopolist produces at a point where the price is greater than the MC. Figure 12 shows the output and the corresponding price under monopoly.

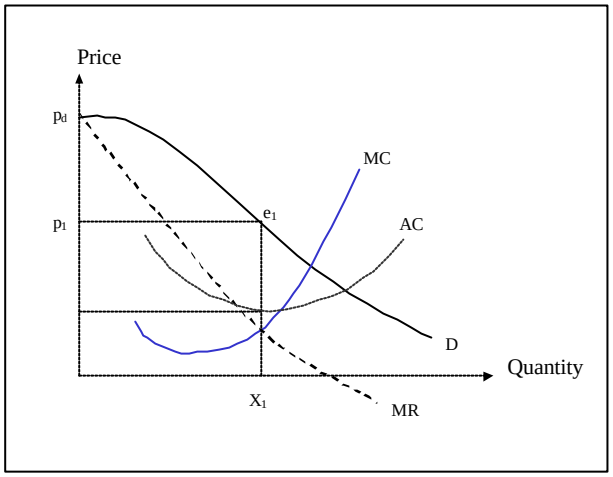

Figure 12. The profit maximization production quantity in a monopoly

Monopolistic Competition

When there are many suppliers in the market, and each of them acts independently in competition with one another, price-making behavior arises when the firms produce imperfect substitutes.
A market characterized by price making monopoly with a large number of suppliers and free entry conditions is called a monopolistic competition. This model of the market is an important tool for analyzing questions of product choices and product variety.

Assumptions

Just as in the case with monopoly, the firms are price makers, which means that they have firm-specific downward sloping demand curves. Neither do they behave strategically and the entry to the market is free. Finally, the buyers are believed to have no influence on the price.

Appropriate Market Structure

There are many buyers and no one of them is large enough to exert any influence on the price. The number of suppliers is also large, which reminds of perfect competition. However, the products/services produced are said to be differentiated and the extent to which customers are informed about prices and available alternatives vary from well informed to poorly informed. At last, no entry barriers to the market exist.

A firm, which is active in a market characterized by monopolistic competition, produces where the MR equals MC. Though, one can observe different competitive market equilib-

rium in the short run and in the long run. With the first alternative, the number of firms is fixed. This case could be treated as a "little" monopoly (see Figure 13). The firm-specific demand depends on the number of rival firms in the market. In the long run there turns out to be a difference between monopolistic competition and monopoly. The reason is that firms will enter the market as long as there are any profits to be made. This will cause firms to produce at a point where *average* cost equals *average* revenue in the long run. However, the condition MC = MR must still be satisfied (see Figure 14).

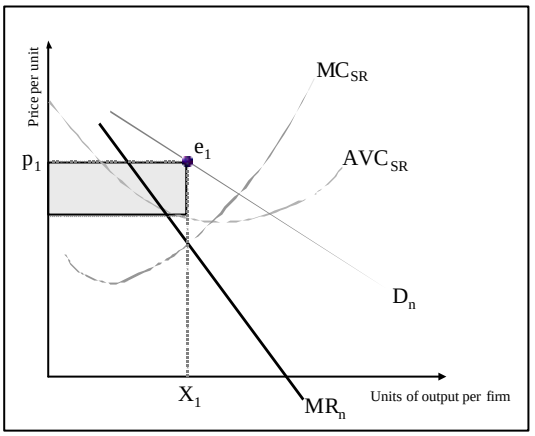

Figure 13. Illustration of the output and the corresponding price under monopolistic competition (short run).

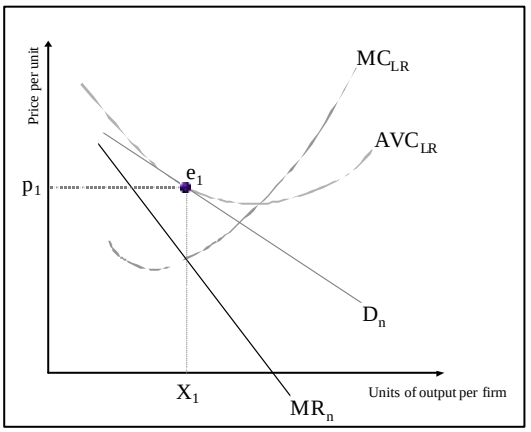

Figure 14. Illustration of the output and the corresponding price under monopolistic competition (long run).

Oligopoly

If firms are aware of the fact that the price or output choice made by anyone of them affects their profits, they are said to recognize their mutual interdependence. Thus, every firm must observe and predict the behavior of competitors. This means that each actor has to behave strategically.

Assumptions

First of all the suppliers in a market characterized by monopoly are price makers. Hence what differs an oligopoly from the previously described market conditions is that sellers behave strategically. The market entry varies from completely blocked to perfectly free and the buyers are believed to be price takers.

Appropriate Market Structure

There are relatively few suppliers but more than one and they sell products, which could be either perfect substitutes or completely differentiated. The extent to which users are informed about the market prices and the available alternatives encompasses both well-informed and poorly informed customers.

Quantity-Setting Oligopoly

In order to describe the market characterized an oligopoly, it could be advantageous to restrict the model to two firms, due to the complexity. Consequently the possibilities to enter the market are completely blocked. Besides the firms are assumed to produce homogenous products and constant marginal costs. These assumptions make the outcome easier to analyze.

Each firm's profits depends not only on the firm's own output, but also on its rival's output level. Therefore, firm A chooses its course of action depending on what it believes B is doing. This is often called a *best response*. The market is in equilibrium when every firm in the market follows the strategy that is a best response to the strategies of other firms.

Assume for example that a firm α sells X subscriptions of a particular service at a price of x SEK/subscription. Given that α sells X subscriptions, β's profits is maximized by selling Y subscriptions. This self-enforcing agreement results in an equilibrium called *Cournot equilibrium.*

Appendix B: Abbreviations

ABR	Available Bit Rate
AOL	American Online
ATM	Asynchronous Transfer Mode
B2C	Business to Consumer
BSS	Base Station Sub-system
CN	Core Network
DiffServ	Differentiated Services
EDGE	Enhanced Data rates for GSM Evolution
ETSI	European telecommunications Standards Institute
FTP	File Transfer Protocol
GGSN	Gateway GPRS Support Node
3GPP	3rd Generation Partnership Project
GPRS	General Packet Radio Service
GSM	Global System for Global Communication
HSCSD	High Speed Circuit Switched Data
IBP	Internet Backbone Provider
IETF	Internet Engineering Task Force
INDEX	The Internet Demand Experiment
IntServ	Integrated Services
ISDN	Integrated Services Digital Network
ITU	The International Telecommunications Union
IP	Internet Protocol
M3I	Market Managed Multi-Service Internet
MPEG	Moving Picture Experts Group
MSC	Mobile Switching Center
MSP	Mobile Service Provider
MT	Mobile Terminal
MVO	Mobile Virtual Operator
NO	Network Operator
PAL	Phase Alternating Line
POTS	Plain Old telephone System
PSTN	Public Switched Telephone Network
QoS	Quality of Service
RSVP	Resource Reservation Protocol
RTP	Real Time Protocol
SGSN	Service GPRS Support Node
SLA	Service Level Agreement
SMS	Short Message Service
TCP	Transmission Control Protocol
UBR	Unspecified Bit Rate
UDP	User Datagram Protocol
UMTS	Universal Mobile Telecommunication System
UTRAN	UMTS Terrestrial Radio Access Network
VOIP	Voice Over IP

VPN	Virtual Private Network
WAP	Wireless Application Protocol
VASP	Value Added Service Provider
WCDMA	Wideband Code Division Multiple Access