

# 1

## Introduction

John Cullen, Mattias Wahlqvist and Gerardo Gómez

### 1.1 Mobile Services in Perspective

Twenty years ago mobile phones were a rarity with less than 5 million subscribers worldwide. They tended to be fitted to cars as car phones as they were bulky and power hungry, used by the elite due to the high prices charged for equipment and service, provided only voice call capabilities and only delivered service over what we would consider a small area today. At the same time, even those companies launching mobile services predicted that the overall market would be very small. Ten years later, many industry observers still believed that the market would remain relatively small.

Today, mobile devices are used by around 1.5 billion people worldwide, a three-hundred-fold increase since 1985, which equates to a worldwide penetration slightly over 20%. Mobile communications is now a technology for everyone. For many people it is now an indispensable part of their life with their mobile being among their key personal possessions alongside their watch and wallet.

The mobile device has changed all our lives and the way we live it. Listed below are a number of examples.

- *Mobility*: Today, we are travelling more for both business and leisure. This has led to a heavier reliance on the mobile phone to stay in touch with colleagues and friends/family.
- *Planning*: As everyone is reachable, we do not plan ahead. How many times have you heard: 'Yes. Let's meet at 12 in the city centre. I'll call you when I arrive, so that we find each other.'
- *Communities*: Part of the tremendous popularity of mobile devices is that wherever you are, communication-wise you are very close to your friends and colleagues. Teenagers today rely heavily on their mobiles to keep in touch with their friends and to organize their social lives. To do this they heavily rely on text to communicate with their community.

- *Participation TV*: TV shows are trying to appeal more to their audiences by allowing their audiences to interact with their shows so as to affect the outcome of the show (e.g. reality TV shows) or generate content (dating/chat shows) while also providing a revenue generation opportunity.
- *Marketing*: Many consumer brands have started launching competitions or offers whereby entries are made via SMS and an instant response can be given to customers. In some cases, prizes are downloads for handsets that allow customers to personalize their handsets with ring tones or wallpapers. At the same time, the consumer brands are able to build up marketing databases using entry information.
- *Security*: Today, most of us would not think of travelling long distances in a car without having a mobile with us in case of an emergency. Also today, in richer countries, many parents are giving their young children mobile phones so that their children can contact them in an emergency and so that they can keep track of their children.

For many young people today, their first commercial relationship with any communications company is with their mobile operator. For this wireless generation, the mobile is at the hub of their social lives. When they first move away from home, they maintain a relationship with their mobile and in most cases the mobile service becomes the only communications service they subscribe to themselves. As a consequence, their relationship with a mobile operator is their prime relationship with the communications industry replacing the traditional prime relationship enjoyed by fixed operators. Today, this unique relationship with the mobile industry tends to be broken only when an individual moves into their own property and starts to consume services that require fixed lines such as broadband Internet services.

Looking forward, we are setting out as an industry on a new phase of market development where with Third Generation (3G) radio technologies the number of services and the richness of those services is greatly expanded. Five years ago, the mobile industry talked about the highest data rates that would be available for 3G. These high data rates are still an issue for laptop PC users with data cards, but for average handset users 3G opens up the ability to use new richer services and capacity that would not have been possible for mass-market customers with Second Generation (2G) technologies. Listed below are a number of examples of how mobiles could be changing our lives in the future.

- *Communication*: Video calling is starting to allow consumers to communicate face to face and to share their environment with their colleagues. In today's busy world with frequent travel, it allows families to keep in touch while on the move.
- *Entertainment*: Music download and streaming is allowing people to get and listen to music on the move, releasing them from computers and fixed communications. At the same time, the ability to download games, which is possible today, will be enhanced by the capability to play them on the move with online friends so providing a new dimension to gaming.
- *Current affairs*: Already the first 3G operators are offering consumers the ability to keep up with events on the move via video clips so allowing consumers to be able to see, for example, their team winning a sports game while on the move.
- *Content creation*: The emergence of smart phones incorporating cameras, good quality displays and reasonable processing capabilities will allow consumers to create and share content. Content could be owner-generated pictures, videos, audio, text or any combination

of these media types. Sharing could be by picture/video messaging, via online electronic journals (blogs) or by peer-to-peer file sharing. To safeguard personal content, network backup capabilities will become essential.

- *Purchasing*: The arrival of large colour displays on devices will make it more practical for consumers to buy services from the Internet and carry out transactions on the move so freeing consumers from their fixed PCs and allowing them to make use of dead time when travelling, waiting for friends, etc. It will also provide a rich channel for governments to communicate and interact with their citizens.
- *Business*: On our company networks today we have from our PCs high-speed access to company resources and to the Internet. On the move, our PC connectivity has been limited by either connection speeds or the availability of hotspot coverage. The emergence of 3G technologies will enable us to improve this situation by providing coverage over large geographic areas.

Like the Internet world, the success of mobile data services will be built for giving consumers access to a rich set of services so as to satisfy a multitude of customer needs at the right cost. Unlike the Internet world, in the mobile environment the winning services and service providers will be determined not only by the simplicity of using services on the move but also by the quality of the experience in using services – the best service in the world will not sell if a user needs an answer in ten seconds and it takes one hour. This book aims to look at how the service performance can be tailored to give the right performance at the right price.

## 1.2 Mobile Technology Evolution

Today, mobile telephony is a global industry with a global footprint in a large part of the populated world. In the beginning, however, mobile telephony systems were typically a local solution on country level.

### 1.2.1 Reasons for Mobile Technology Evolution

There has been a tremendous evolution of mobile telephony during the last 20 years, both technology-wise and service-wise. One interesting aspect of the evolution of mobile technologies is to ask yourself what is really the driving force being the engine for the switchover from one technology to another. That is a complex question, and there is not one true answer. It is also so that the answer will depend on whom you ask. Here we anyhow try to illustrate the complexity of this question by giving a few opinions from different points of view.

- *Customer service requirements*: Is it so that end-users are demanding better and more requiring services, which leads operators and vendors to implement new technologies? This statement is partly true and it is important to observe that it will likely become truer as time passes. In the beginning, the mobile telephone service was just a telephone service you could use on the go. Today, there are additional services (SMS, WAP etc.) that are adding new requirements to the system. It is also so that the end-users today are much more advanced in terms of comparisons with, for example, services on the fixed Internet.

If a person can download a large email on the fixed computer, there should be no reason why he/she should not be able to do it in his/her mobile phone.

- *Customer and traffic growth*: Is it so that the growth in the customer base and the traffic that generates are implying that the operators need to reinvest in newer more efficient systems? This is not really true. Typically, new features (e.g. half-rate codecs, frequency hopping etc.) are introduced to enhance capacity and quality, but it is of course important for the operator to protect his CAPEX investment as long as possible. It is also so that the time to design a new system makes it impossible to rely on a new more spectrum efficient handling of the traffic. The problems are here today, and the future system will take many years to get into the field.
- *Differentiation of services and Quality of Service (QoS)*: Is it so that new systems are developed to be able to perform service differentiation and offer QoS? To some extent yes. It is a common understanding that service differentiation and QoS is the only way to cost-efficiently offer a wide range of services. Still, the service differentiation has already been gradually introduced in today's systems, and so making service differentiation a main reason for the development of new systems is only partly true.
- *Spectrum availability*: When new spectrum is made available there is of course an urge to make use of it in the best possible way. Spectrum is a scarce natural resource, and the introduction of new more efficient systems is done easily if it is introduced together with a new spectrum band.
- *End-user requirements*: The end customer has normally a firm opinion on whether he likes a service or not ('like' in this context normally means that he thinks that it is worth paying the stipulated price for getting the service). That opinion heavily affects his usage of the service. Still, considering the time it takes for a service to become a mass-market service, makes us believe that it is not end-user requirements that are driving the need for new systems. The majority of the end-users are not advanced enough to know what they will need in a five-year time frame.
- *Commercial aspects*: There are of course commercial aspects that influence the willingness to introduce new systems into the markets. Vendors might want to protect or increase their market share; operators might want to create a high-end profile towards their end customers etc. Considering the time frame to introduce new systems, it is anyhow clear that the commercial aspects are mainly considered on high strategic level.

To conclude, we can see that there is a variety of reasons for new mobile systems to be introduced, with the strongest ones being the need to make more and more efficient use of a limited natural resource, the spectrum. On top of that, there are a multitude of other reasons to consider.

### 1.2.2 Mobile Technology Evolution Paths

Analog technologies were dominant in the cellular market up to 1997, when their global market share was exceeded by that of 2G digital technologies. From that date, the Global System for Mobile communication (GSM) revolutionarily changed the way we look at and think of mobile telephony. After its introduction we have seen a rapid evolution of services, technologies and performance. GSM technology's market share shows a sustained growth and today it has become the global 2G standard, deployed by more than 460 operators around the world and accounting for more than 70% of the total number of cellular subscribers.

General Packet Radio Service (GPRS) technology, developed as a Packet Switched (PS) extension of the GSM network, allowed high-speed access to IP-based services and at the same time it provided an efficient use of the network resources. Some time later, Enhanced Data for Global Evolution (EDGE) technology increased the radio data rates by including some enhancements in the modulation and coding schemes. (E)GPRS can be considered as the convergence point between the Time Division Multiple Access (TDMA) developed in North America and GSM technologies, and is the foundation for the PS domain of the 3G Universal Mobile Telecommunications System (UMTS).

Another parallel technology evolution path is the one coming from cdmaOne. Despite an important growth during its first year of deployment, cdmaOne's (and its main successors: CDMA2000-based family) market share has stabilized around 15% of market share in 2004. Although a natural evolution from CDMA2000-1x would be the support of 1xEV-DO (1xEvolution, Data Optimized) and 1xEV-DV (1xEvolution, Data and Voice), many CDMA operators are currently migrating towards GPRS and EDGE technologies as an alternative option (with the later integration of WCDMA). This last option is however dependent on the cellular-operator's licensed bandwidth, since WCDMA technology is currently not supported in the 800-MHz band, the future availability of dual mode cdmaOne/WCDMA terminals and the integration effectiveness of the technologies.

Expected WCDMA launch is nowadays becoming a reality as the evolution path of 2G technologies, being already supported over several markets around the world. The convergence of 2G technologies towards the UMTS multi-radio 3G evolution path is clear. The entities, such as operators, global associations and standardization bodies, which are representing and driving the evolution of three out of the four current most representative 2G technologies, have endorsed the UMTS multi-radio evolution path. Figure 1.1 summarizes the evolution paths associated with the existing 2G technologies.

The evolution of the mobile technology's market share and how these technologies are distributed around the world is depicted in Figures 1.2 and 1.3, respectively [1].

Thanks to the evolution of the networks towards PS technologies, data services have experienced a huge increase in terms of data transmission capabilities, leading to an important increment in operator revenues. Currently, SMS and MMS are still the most profitable,

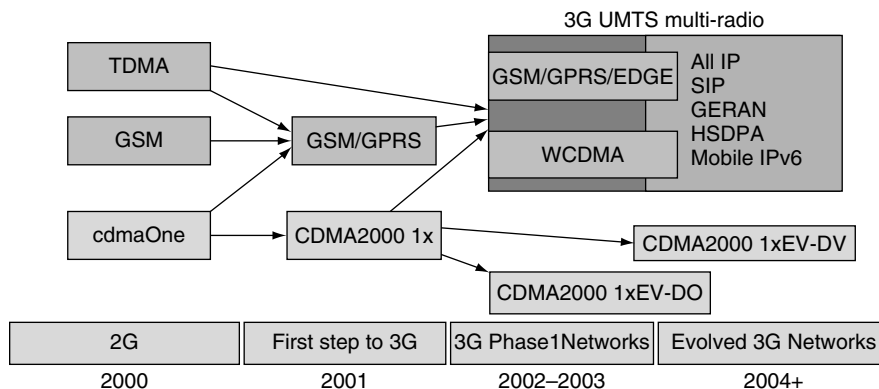


Figure 1.1 Mobile technology evolution paths [2]

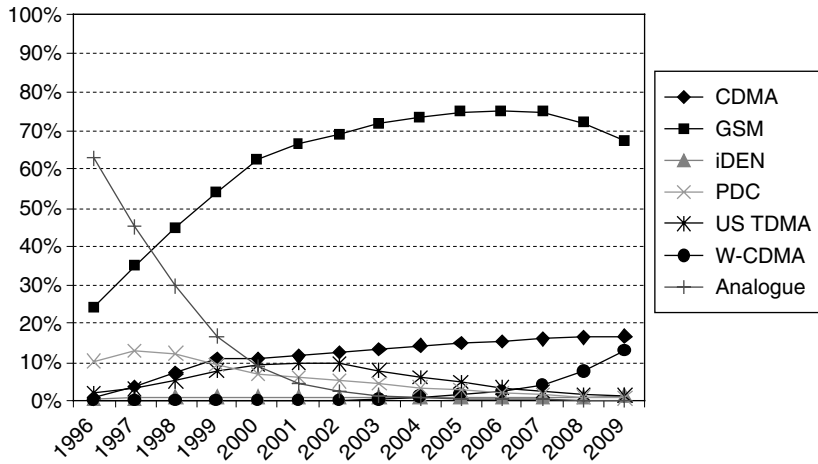


Figure 1.2 Mobile technology’s market share (forecasted from 2005 onwards) [1]

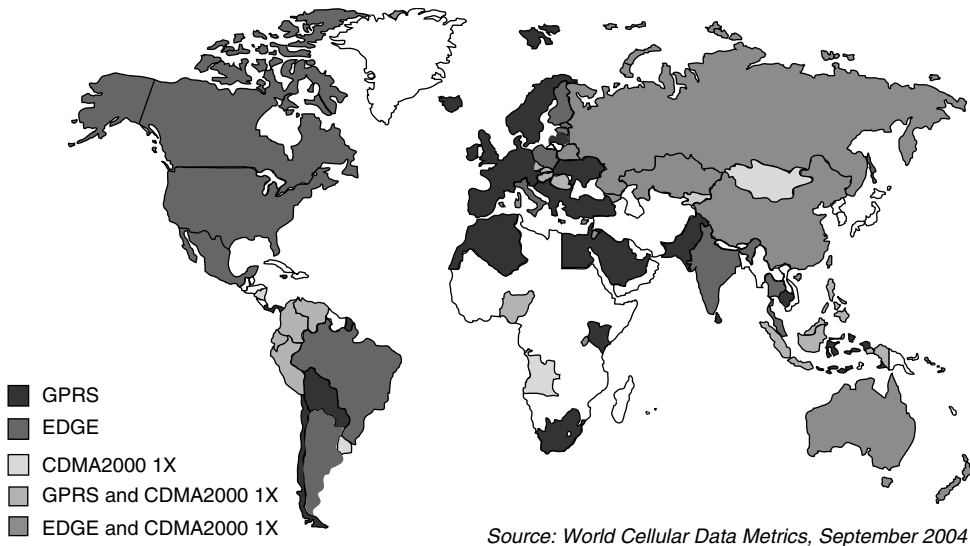


Figure 1.3 Mobile technology deployment [1]

although other services like email, content downloading (i.e. Java applications, games, tones, etc.) or streaming are already pushing hard.

SMS and MMS, together with ring tones and information downloads, have represented between 2 and 7% of operator’s revenue in both North American and Latin American regions in Q2 2004. China Mobile handled more SMS than any other operator, 30.9 billion in Q2 2004 [1]. Although in those regions, CDMA2000–1x was the most widely deployed technology,

data usage has been boosted by the continued deployment of advanced data networks (GPRS/EDGE), as shown in Figure 1.3.

In Europe, GPRS and partially WCDMA have been deployed until today, where an average data percentage of revenue reached 13.6% in Q2 2004. SMS traffic in western Europe grew approximately 17–18% in the 12 months by the end of June 2004 [1].

MMS had been launched commercially by 237 operators in 88 countries in September 2004. MMS usage and traffic volumes on the whole remain low, being KTF Korea and Verizon (USA) the ones reporting a higher number of MMS (over 21 million in Q2 2004).

Total mobile subscribers to GPRS, CDMA2000-1x, I-mode and other advanced data services exceeded the 150 million mark in Q2 2004, and the total reached just over 152 million as at 30 June 2004, or 9.9% of the world's total mobile users. The reader is kindly referred to Chapter 2 for a detailed description of the different technologies listed along this section.

### 1.2.3 Harmonization/Evolution Challenges

With the design and commercialization of a new system, there are a large amount of requirements that need to be considered not only from a technical perspective, but also from an economical and commercial point of view. Here we list a few major challenges that are important to consider.

- *Backwards compatibility:* In order to get maximum reuse of older investments, a natural evolution also requires the new system to be backwards compatible towards older, already commercially deployed systems. For example, with the introduction of UMTS in Europe, it is of high importance that already from day 1 being able to perform inter-system handovers to and from the commercial deployed GSM network. The reason for this is obvious, as the operator wants to be able to offer continuous service coverage. Note, however, that general requirements on backwards compatibility can mean many very different technical requirements on the new system. An operator might want to make maximum use of his already deployed network, which might lead to, for example, requirements on inter-system handovers and the ability to co-site the two systems. On the other hand, a mobile phone vendor might want to ease the implementation of a multi-system handset, and might want to set requirements on clock frequencies to be used, as well as limiting the complexity between how the two systems interact.
- *Service transparency:* New systems typically offer new services. For that reason it can be difficult, especially in the beginning, for the operator to offer a continuous service support over the whole coverage area. For example, high-speed data service is impossible to maintain in UMTS when the user leaves the UMTS coverage area and are handed over to GPRS. From this aspect, it is anyhow considered important to be able to maintain some type of service, even if the service level is lower. Whether that is useful or not for the end-user is anyhow very service dependent.
- *Interoperability/roaming:* As users move between systems, it is difficult for the operator to maintain a constant service level. In the same case as in the service transparency example above, a user that is roaming into another network might not get all the services he/she can normally access in his/her home network, as they might simply not be implemented. Another consequence is the implications this might have on the billing models to be used.

### 1.2.4 Future Outlook

It is clear that the development will not stop here and now. New systems and features are today being standardized and developed for inclusion in the upcoming years. A global traffic growth together with the release of new spectrum or reforming of old spectrum will also increase the need for newer more efficient ways to transport mobile data communication. Lately, a few trends have however emerged that might have significant impact on how future mobile communication systems are designed and deployed, although they are not going to change the evolution path.

- *Emerging markets:* There are still large parts of the world where mobile communication has not yet been deployed and there is a large market potential. Typical situation for these countries is that they have a high potential subscriber base, and that the fixed phone infrastructure is not so developed. What holds back a massive deployment in these countries is normally that the average amount that can be charged to the subscribers is relatively low, which makes it challenging to deploy and market a network in a cost-productive way. Ultimately, this might lead to requirements to develop 'low-end' systems with lower production cost and less features.
- *New services:* With the exponential growth of the Internet, there is also an explosive amount of new services that users can access. By getting used to access these services on the fixed Internet, there will be a demand for doing at least some of them while being mobile. In addition, mobile communication can offer a multitude of mobile services that will also add new requirements on any future system.
- *New users:* The introduction of mobile data services also opens up for a complete new subscriber group – machines. The vending machine can itself send its order for new drinks or to get service, or you could remotely find your car on a map and demobilize it when it gets stolen. This is an area becoming more and more important with an infinite amount of possibilities that we likely will see.

## 1.3 Motivation for QoS

The motivation to look at QoS is two-fold.

1. To provide a service experience to consumers that meets their expectations so that they are more likely to use it again and recommend it to friends or colleagues.
2. To achieve optimum loading of an operator's network so that the desired service experience is delivered for each customer while maximizing network utilization.

The following section provides a brief introduction to the main factors involved in addressing these two issues.

### 1.3.1 Service Experience

In the early data services market, many consumers are impressed to just use a service when mobile. However, once this euphoria has passed, the vast majority of consumers start to judge a service based on how it meets their needs and expectations. As an example, the Short

Message Service (SMS) was designed as a store and forward service, and was offered as such by most operators in the mid-1990s with messages sometimes delayed hours before delivery. Consumers have conclusively taken to the service as a way of communicating quickly and efficiently with friends and colleagues without the need for a conversation. However, in meeting these needs we expect that a message is delivered almost instantly. When important business and social meetings are arranged by SMS, a couple of hours delivery delay is unacceptable. As we can see from this example, a consumer's expectations of a service dictate whether it is perceived as working well or badly. These expectations in turn determine the critical success factors that the network must deliver against if the service is to be perceived as good. Take the following examples.

1. A customer using an 'always on' email application (e.g. Blackberry) expects their emails to be accurately received and that they are received within a reasonable time, for example 10–20 minutes after being sent. This implies that the network must deliver accurate information, i.e. a very low bit error rate, but that the payload can be delayed for a reasonable amount of time.
2. A customer using a Push-to-Talk application will expect to get voice messages within a couple of seconds from their friend or colleague sending a voice message but must be prepared to tolerate some voice distortion on limited occasions. This implies that the network must expedite the voice messages through the network but that limited packet loss can be tolerated.
3. A customer browsing the Internet from a laptop PC with a 3G data card will expect that the Web page loads accurately to a point where they can start reading in less than 10 seconds, otherwise their concentration will lapse making the service uncomfortable to use. This implies that the network must deliver accurate information and that some limited delay can be tolerated. In this example, the way the Web page is built can also make a difference. For example, a Website that displays text within 10 seconds but then downloads images in the next 10 seconds will often appear to be quicker than a page that completely downloads in only 15 seconds.

In all these examples, if the network achieves the critical success factors, the consumer is likely to perceive a service as working well. If the network fails to deliver, the consumer is likely to perceive a service as working poorly.

When considering the quality of a network it is worth remembering that services run either between two terminals or a terminal and a server, and that the critical success factors apply across the whole connection. There is no point, for example, in engineering the GPRS network to meet the consumer's expectations when the connection to the content provider is not to the same standard and hence degrades the overall experience.

The main causes of a network failing to deliver against the critical success factors are:

- *Radio network performance* – Are there a lot of errors on the radio interface?
- *Network capacity* – Is there sufficient capacity to deliver a good service?
- *Network design* – Is there too much delay in the system; is sufficient capacity available end to end?
- *Application design* – Are the right protocols being used for a mobile environment?
- *Service support* – Is service enhancement technology correctly configured?

### 1.3.2 Radio Network Performance

A well-planned radio network where data errors on the air interface are minimized in most cases will improve application performance. If there are a large amount of errors, retransmissions are required which can slow down the amount of information that can be transferred by protocols such as TCP. Radio errors also introduce extra delay into any conversation between application clients/servers, slowing down application response times. In a similar way, voice conversations can be slowed down by satellite delays.

### 1.3.3 Network Capacity

Well-designed mobile networks are dimensioned so that they have just enough capacity during busy periods. Any more capacity than necessary adds network cost for operators reducing profit margins. Too little capacity and customer applications will not be able to get network capacity so will deliver a poor customer experience. It is, however, extremely difficult to predict loading accurately as demand may fluctuate by time, day, month and season as well as demand growing with time. As a result, any operator wishing to offer customers a good experience, so as to encourage the uptake of services, would have to expensively over dimension their network to avoid congestion. To overcome this problem, standards have defined the concept of QoS and this is starting to be implemented into network equipment. The QoS concept encapsulates the idea that different data streams could be treated differently by the network depending upon the service being carried (Chapter 4). Ideally, a service that requires fast response time is assigned a QoS that in periods of congestion it would receive priority over other traffic. Conversely, a service that can tolerate a reasonable delay would have lower priority than other traffic. By assigning different QoS to different services, when congestion occurs traffic can effectively be smoothed over time with high priority traffic still being transmitted with little delay but lower priority traffic being delayed until capacity is available. As a result, in periods of temporary congestion, a network providing QoS can meet customer-service expectations with existing capacity. In this way, with QoS, operators can more effectively load their networks as they can tolerate temporary congestion while at the same time ensuring that they deliver customers with the good service experience they expect.

It is worth noting, however, that QoS mechanisms are designed to work in periods of temporary congestion, where lower priority traffic can be delayed without impacting the service experience of those services. If heavy congestion occurs or the network is congested for extended periods of time, QoS cannot be relied on to maintain the customer-service experience. In these cases, further capacity is required and the network should be re-dimensioned to a level where only temporary congestion occurs.

### 1.3.4 Network Design

The principle aspects of network design that can impact service performance are:

- *System delays* – Every additional piece of end-to-end delay slows down application ‘conversations’. For services where customers demand low response times, these delays can make a service unusable if the application requires an extensive ‘conversation’.

- *System design* – An operator may choose to set up a data session when the phone is switched on, requiring costly resources to be allocated as long as the phone is on, or when a service is activated, adding additional delay while the data session is set up.
- *Equipment* – One node will always be the system bottleneck. It is important to understand where bottlenecks may develop and understand the scenarios in which they will occur. For example, a router could be limited at any point in time by the number of active sessions, the data throughput per second or the number of packets it can transfer in a second. In this example, changing the mix of services in the network from predominant Web browsing to Voice over IP could change the place and type of bottleneck.

More information on the impact of service performance on design can be found in Chapter 5.

### 1.3.5 Application Design

Application design can impact on service performance in two ways. First, the user-interface can be designed so as to soften the impact of network performance on the user, as described in the Web page example earlier. Secondly, the application can be designed so as to better carry out communications over wireless connections. This book leaves the issue of user-interface design to usability experts but instead chooses to focus on application design. In focusing on application design, particular attention should be paid to:

- *Protocols used* – TCP assumes that packets that do not arrive in a certain time period are lost due to congestion and therefore the amount of information that can be transmitted at any one time should be limited. In radio networks, these ‘lost packets’ may have nothing to do with congestion but could be the result of a temporary radio failure lasting milliseconds. As a result, protocols used to build new services should be considered carefully.
- *‘Conversation’ structure* – In order to deliver a service, application clients and servers tend to have a dialogue so as to transfer information on the service requested, the application/person requesting, what format the service should be delivered in, etc. If the dialogue is sequential, one bit of information is transferred and acknowledged before the next bit can be transferred and acknowledged. In mobile networks with high delays the overall effect of the delays is multiplicative. This leads to unacceptable service delays from a customer perspective. Ideally, for good performance as many actions are performed in parallel as possible. One consequence of this is that when using HTTP, which provides transport for Web page requests and responses, version 1.1 considerably outperforms, by up to a factor of 3 times, version 1.0 simply because it allows parallel processing of Web page components.

More information about impact and interactions between the wireless system and different applications can be found in Chapters 3 and 5.

### 1.3.6 Service-Enhancing Technology

The aim of service-enhancing technologies is to improve the performance of data services by overcoming limitations in wireless systems caused by the radio environment. Examples of service-enhancing technology are:

- *Payload compression* – Information transferred across the radio interface is compressed first so that it takes less time to transfer over a radio connection.
- *Controlled quality degradation* – Higher quality and resolution especially of pictures implies larger file sizes and thus longer transmission times. When the final end of the transmission is a mobile terminal with limited screen resolution, in addition to normal compression, images can be downgraded or reduced in size so that they are still recognizable and fit better to terminal screens. This process is a compression with losses and can perform in network elements such as proxies.
- *Proxies* – Proxies store copies of Web page components locally so that when a Web page is requested, the proxy can instantly provide the Web page components without incurring further delay as it avoids having to get the Web page from the Internet. As Web pages can be made up of many components which need serial requesting, cutting the end-to-end delay associated with collecting components can have a significant impact on Web page download times.
- *Protocol optimization* – Protocols are adapted to make them more suitable for use in the radio environment. The most common optimization is related to the TCP protocol which is used to carry the majority of Web traffic. The optimization aims to counteract the reduction in flow rate associated with delays to TCP acknowledgements, which TCP interprets as congestion, by injecting fake acknowledgements into data streams. At the application level, optimization can take the form of consolidating all the Web components into one transaction, which is downloaded so that long delays caused by requesting and receiving Web page components serially is avoided.

Service-enhancing technologies are usually implemented together with performance enhancing proxies, which are treated with more details in Chapter 8.

### 1.3.7 Conclusion

The aim of service performance optimization is to provide a service experience for customers that meet their expectations while maximizing network utilization. Providing a good service performance requires operators to examine and optimize how they deliver services at multiple network and system levels. Today, many operators spend significant effort in optimizing their voice networks. Optimizing service performance for data services is likely to be a significantly more complex and resource-consuming task as there are considerably more variables. The following chapters of this book will examine each of the issues raised in this section in considerably more detail so as to provide guidance on how to achieve the best service performance in mobile networks.

## References

- [1] EMC Database: [www.emc-database.com](http://www.emc-database.com).
- [2] T. Halonen, J. Romero and J. Melero, 'GSM, GPRS and EDGE performance', Ed. Wiley, 2nd edition, 2003.