Chapter 8 MOBILITY IN CELLULAR NETWORKS

This chapter will provide a complete overview of the mobility models used in cellular wireless systems and the location management techniques as the two main parts of mobility management required in mobile networks. In a wireless network, the users are assumed basically to be mobile, which means that they will change their network point of attachment frequently, irrespective of whether they are idle or active in terms of exchanging data with the network and other network users. Therefore, understanding the user mobility models and location management techniques becomes profoundly important.

Chapter 9 TRANSPORT PROTOCOLS FOR WIRELESS IP

Transmission control protocol (TCP) is the *de facto* transport protocol for today's global Internet. It performs at an acceptable efficiency over traditional wired networks in which packet losses are usually caused by network congestion. However, in networks with wireless links in addition to wired segments, this assumption would be insufficient, as the high wireless bit error rate could become the dominant cause of packet loss, thus making TCP perform suboptimally under these new conditions. In this chapter, the suitability of TCP for wireless IP will be discussed and several techniques that could improve the situation will be explained.

Chapter 10 INTERNET PROTOCOL FOR WIRELESS IP

Network layer has a distinguished role in the realization of future wireless IP networks. The network layer functionality could determine the efficiency of the wireless system and its performance in terms of quality-of-service parameters. The main part of the network layer is the network protocol. In today's Internet, it is the Internet protocol that functions as the network protocol. Therefore, for the study of future wireless Internet networks, an in-depth understanding of Internet protocol and recognition of its merits and drawbacks is of vital importance. This understanding will assist researchers in the field to find solutions for improving the Internet protocol toward its position within the future wireless IP networks. This chapter will provide such an understanding by reviewing the current version of the Internet protocol, its next-generation version and also the initiatives toward migrating the Internet into the mobile environment.

The Wireless Mobile Internet
Abbas Jamalipour
Wiley 2003

5 Quality of Service in a Mobile Environment

Quality of Service (QoS) is a terminology that has been around for quite some time and maybe it is one of the most common words that have been used by the people in the field of telecommunications in the past few years. The popularity of this usage, however, could not provide a clear definition of this important term and in many situations leads to misunderstanding and confusion. When the meaning of the term is not clear, then it would be difficult to provide its capabilities in a system, namely, a QoS network, as the expectations from different people would be different.

In this chapter, we try to define the QoS very clearly so that we know for sure what the problem is and how we can establish a network with guaranteed quality in its services. Our discussion on QoS will be more general than QoS being only used for wireless Internet, that is, the subject of this book, but we will try to limit the discussion to the usage and deployment of QoS in wireless Internet. After defining the QoS term and its requirements, we will establish a framework for a QoS network and exercise the Internet protocol (IP) and wireless cellular initiations for the QoS. Finally, we will run through wireless data networks and wireless Internet to see the prospects of QoS in these emerging networks.

5.1 DEFINING THE QUALITY OF SERVICE

QoS can be defined as a set of specific requirements for a particular service provided by a network to users. These requirements, however, are usually described by using some quantitative figures. So instead of asking for a good network service, the user is asked to specifically request other sorts of measures such as connection speed or delay, which can

The Wireless Mobile Internet: Architectures, Protocols, and Services. Abbas Jamalipour
 ISBN: 0-470-84468-X

be described by a numerical value, for example, 56 Kbps or 300 ms, respectively, for the speed and delay. Having a quality term such as good or bad described by a quantitative metric simplifies the process of allocation of that quality to a particular service by the provider and also prevents any possible ambiguity during the user request and service fulfillment process.

Changing quality into a measurable quantity is a good step toward clearing the meaning of the term but still there are other things that have to be done. The ambiguity in QoS definition can be pointed out in many directions by posing some of the following questions:

- What types of qualities can be attributed to a particular service?
- Which entity in the network is responsible for providing a particular service?
- Who is the accounting entity in the network that authorizes a user to receive a particular service?
- Is the requested service a network service or an end-user terminal service?
- Who judges the need for a particular service, the user or the network?
- What is the source of the service requirements: user, network, or technology?
- Is it possible to fulfill a service quality at a particular time and location?
- What should be done if the requested service is not available at a particular time or location?

The above questions and many similar ones simply illustrate that there cannot be a definite definition for the QoS and any attempt to find a single answer would be a waste of time. In recent years, many companies have started to raise the issue that the next-generation telecommunications networks, including Internet users, can receive a guaranteed QoS. However, the topic is too broad and no one can explain what the guaranteed QoS would be. Everyone knows that if a network can provide a full commitment on its services at all times, then it would be the ideal network, but the issue is what type of services we are considering, are they fixed for all users or different on a user base. We will explain this issue in this section by separating user-based, network-based and technology-based requirements of the QoS. The reader may also refer to some survey papers on QoS for further reading [1–4].

5.1.1 User-level QoS requirements

At the user level, requirements of QoS are mainly those that can be seen by the user. This means that many system-level QoS architectures are essentially transparent to the user or that because they are not directly related to the end-user service, they are ignored by the user. For example, if you are using a cellular phone for your daily voice communications, your ultimate purpose and requirement from your service is to have a reliable phone conversation and a reasonably acceptable level of voice quality. If you move from the service area of one base station to another one, the cellular network needs to perform a complicated procedure of location updating and handoff between cells in order to maintain the continuity of your call and service. For the user, all these procedures are transparent

and ignored, unless a call drop or sensible change in quality of communication happens during the handoff.

The above example raises an important parameter in defining the QoS at the user level. This is the user application. In the above example, a voice application has been used and on the basis of that application, the user expectations of communications continuity and voice quality are considered as the perceived QoS metrics. We can generalize this example to any type of network including cellular, Internet, wireless LAN, and so on. So we conclude that from the user point of view the user application has a determining role in defining the QoS requirement.

In general, the user-level QoS requirements can be categorized into three:

- Criticality
- Cost
- Security

Criticality is defined in accordance with the perceived QoS based on data transmission and application type. To make this clear, let us consider an example of multimedia transmission of a video clip over the Internet with voice. A user may consider several factors in order to illustrate either satisfaction or a problem with such a communication system. These are:

- Video rate
- Video smoothness
- Picture detail
- Picture color accuracy
- Audio quality
- Video/audio synchronization

The topic of multimedia QoS is discussed in many literatures (e.g. see References [5–9]). These factors are self-explaining and are usual in the case of videoviewing on the Internet. But the important thing that is not very visible and that we want to conclude from this example is that not all users use the same set of factors to determine the quality of such a video transmission. One might be interested in smooth playback of the video at an acceptable rate without paying much attention to the audio quality or even the picture color quality or picture detail. Therefore, it is not the case that even for the same application we put the same set of QoS requirements for all users. Users can choose the application and there are a limited number of QoS requirements associated with each application (such as the set we listed for the video transmission example). However, among all these requirements there will be a subset that makes the user QoS profile to be satisfied by the network service provider.

The second category in the user QoS requirements is the cost. This is one of the most indicative parameters for users when considering the quality. Cost is the money value of the service fees that the service provider charges a user. When a user wants to include this as a QoS metric, two different types of cost can be considered: per-use cost or per-unit cost. In a data-communication session, it is important for a user to see if the user is

charged either on the basis of the period of communications, that is, the usage time of the service (e.g. measured in seconds), or according to the amount of data (e.g. measured in bits) transmitted in the uplink and the downlink from and to the user. Charging on the basis of the amount of data has become very interesting for many users recently, thanks to the increase in Internet and data applications compared to traditional voice and other constant rate and real-time applications. This can be considered an important factor in the success of i-mode for wireless Internet.

The third category for QoS requirement is security. This can be further divided into several types, such as

- Confidentiality
- Integrity
- Digital signature capability
- Authentication

Confidentiality refers to the service in which a particular user's information can be accessed only by appropriate and recognized users. If the user is worried that the information is not to be corrupted during the course of transmission, then the integrity becomes the desired security QoS metric. Digital signature provides an example of methods by which a user can make sure of the identity of the sender of a particular package of information. So if you are among those people who do not open a mail when there is no sender address on it, then for an electronic mail session, you will need this type of digital signature too. They also provide a safe method of making sensitive financial transactions. Verification of a user's identity and the right to access a particular service and information is provided through a process called *authentication*.

As we can see, for the second and third categories of the QoS requirement, there still is a general set (maybe expandable from what we listed here) and the user has the freedom to choose and customize his own subset. The chosen subset will then be forwarded to the service provider and if both parties agree, then the user QoS profile will be prepared.

5.1.2 Technology and network QoS requirements

On the basis of the technology and the network architecture, we can find more indicative figures to illustrate the QoS provided to users. Although many of these indications can be seen by users, they are more or less related to the technology behind the service and thus a user will find limited flexibility in changing the profile after subscription to the service.

We may categorize these requirements mainly into three types:

- Bandwidth
- Timeliness
- Reliability

Bandwidth illustrates the speed or data rate available to a user application. Very loosely speaking, we can say that the more the bandwidth available in the system, the higher the

data rate that can be provided to each user application. The statement above is loose in the sense that it does not include other determining parameters that could affect the actual data rate given to a particular application at a specific period of time. For example, when you compare a 100-Base-T with a 10-Base-T LAN, with nominal data rates of 100 Mbps and 10 Mbps, respectively, each with N hosts attached, you may say that the 100-Base-T LAN is faster than the 10-Base-T LAN. However, this comparison does not consider the loading of the two LANs. If, for example, the 100-Base-T network is fully loaded and the 10-Base-T network is very lightly loaded, then each of these N hosts in the latter network can receive a faster connection speed than the users in the former network, because the LAN system is a bandwidth-shared system. For the sake of discussion in this section, we will use this loose relationship between bandwidth and speed. In Chapter 7 that deals with traffic management in wireless IP networks, more discussion will be provided.

To make the definition of bandwidth more precise, we need to distinguish between three different rates, including the system-level data rate, the application-level data rate, and the transaction data rate.

System-level data rate shows the actual data rate at the physical transmission media at uplink (from user terminal to the network) and downlink (from the network to user terminal) directions. This rate relates to static network characteristics such as the type of media used for connecting the user terminal to the network, network technology, and network topology as well as network dynamic characteristics such as current traffic loading, current network capacity, and also the service agreement between the user and the subscribed network. The system-level data rate could have a nominal value that can be within a range agreed to between the user and the network. The system-level data rate could also be further limited by interactions between protocols at the higher layers of the network stack such as transmission control protocol (TCP) and IP.

Application-level data rate could have a completely different value from the system-level data rate. Application protocols designed for high-bandwidth applications such as multimedia usually use different compression algorithms in order to reduce the amount of data exchanged between the application and transport protocols and eventually passed to the physical layer. Usually, the more the application data is compressed, the less the bandwidth that is required to transmit the data, but at the same time this means reducing the quality of the received information. So, there is always a trade-off between how much bandwidth you may use and the quality of information you receive at the other side of the network connection.

The transaction rate illustrates something completely different from either of the above two rates. If you assume a transaction as one single task to be done in the transmission of certain information, the transaction rate simply shows the rate at which the predefined tasks can be performed by the user, successfully supported by the attached network.

Depending on the nature of applications run by the user, one or some of these bandwidth indicators can be used to illustrate the QoS provided by the network. A user might use different indicators, something that is more sensible for human beings such as delay time, to quantify the bandwidth service provided by the network.

The second category of network- and technology-based QoS is timeliness. Timeliness can be sensed through delay time, response time, and delay variation, as well as similar indicators. Delay can be defined as the time spent by a user from the instance the

user requests some information from the network until the instance that the information is completely downloaded to the user terminal. Note that this is not the only way in defining the delay. According to the user or network requirement, you might find another definition more appropriate and useful. However, the definition given here is the most straightforward way for quantifying the QoS in a data network similar to the Internet. Such a delay definition covers all type of delays that could happen within the network, including the user terminal processing time and transmission delay, link propagation delay, queuing delay at the input ports of the intermediate routers, backbone network delays (including link delays and router processing delay), and also the processing delay at the destination host, which has the requested information.

By not using the very loose relationship that exists between the bandwidth (at system level) and delay, we can say that with a higher bandwidth you may experience a smaller delay in most situations. But you need to be very careful here. This relationship always exists but in some occasions it is affected severely by other network parameters so that the concluded result shows an opposite relationship. Since our delay definition includes the processing delay at the destination, for example, it could be the case that even with a high-speed link and network a user experiences long delay when the destination host is too busy or the requested information is accessed by many other users. That is, the traffic load at the destination network or host, as well as loading of the backbone network (partly or entirely), could result in a long delay even with high bandwidth locally.

Response time can be considered as part of the delay definition given earlier or in some occasions as a single indicator of the quality. It simply tries to illustrate how fast the network as a whole is in providing the requested information to a user.

Delay variation is the third timeliness indicator that we have specified. In the usual situation, the previous two timeliness indicators, delay and response time, could be sufficient to illustrate the service timeliness quality. However, for some applications, such as real-time multimedia, it is not the delay but the delay variation that affects the quality. If the network as a whole imposes a long but fixed delay at all times during the period that the application is running (e.g. a videoconferencing), it is possible to compensate the effects of delay by simple methods such as buffering and delayed replay. However, if the delay variation shows very diverse values from time to time, it would be very difficult for the application protocol to adjust to a good method of compensation. Therefore, the delay variation would be necessary in some situations that include increasing usage in future networks.

The third category we have named for the network- and technology-based QoS is reliability. Reliability in a networked system could be quantified by measuring the time or the frequency. The time could show the average time for a failure to happen, the average time the system needs to recover from the failure, or the mean time between failures that happen in the network. It is also possible to count the rate at which system failure, data loss, or data corruption happens in the network.

The reliability measure could be more important for the network than the individual users for quantifying the network QoS. It would be very important for the network to avoid long failure times or very frequent failures or corrupts, as this will affect all users, whereas for individuals who use the network services from time to time, delay and bandwidth metrics will be more visible.

5.1.3 Correlation between the QoS indicators

Until now, we have discussed different QoS requirements and tried to quantify them through numerical measures, so that it could be possible for a user to specify precisely what his expectations are from the attached network in terms of QoS, and also for the network to advertise for its users what type of services can be offered to them. The indicators defined in the previous two subsections are all important and they can define the service quality individually or together according to the application and the users' needs. According to the discussion provided there, however, it should be clear by now that having all those indicators as the QoS metrics for all users is neither feasible nor necessary. On the basis of the user requirements or the application requirements at a particular time, there will only be a subset of those QoS indicators that need to be provided by the network to a user. Therefore, the ambiguity of the QoS should have become clear such that the QoS must be defined on a case-by-case basis and not as a general rule that can cover all requirements.

In addition, we need to note that it is difficult to provide a guideline that can cover the QoS in all situations. We can see a very diverse relationship between the QoS indicators observed in different systems. To illustrate this fact, we use an example from the wireless that is relevant to our topic in this book. We assume that data transmission (as opposed to voice communications) is our required service for the purpose of this example.

For the wireless, we can add another QoS indicator, namely, the mobility range. Mobility range in our definition can have two different meanings. The first one is how big is the geographical area in which a user can move around and can still receive the service from the attached network. Again, similar to the previous indicators defined in previous subsections, this is not something that is independent of the user's application or that can solely determine the QoS for a particular network. For example, if you need a wireless service within the borders of a room, then having a one-mile coverage would not be an advantage in choosing the system to be used. The coverage, however, would be very important for a cellular mobile network in which users want to move very widely.

Mobility range can also be defined as the size of the area covered by a single base station or an access point (AP), for example, the size of a cell in a cellular mobile network. We will now see the relationship between the mobility range (mainly defined as the former one) and other important QoS metrics. Bandwidth (as a network-related QoS requirement) and cost (as an important user-based QoS requirement) are considered here in conjunction with the user mobility range.

Table 5.1 summarizes several known wireless networks and lists their usual coverage (i.e. the mobility range offered to the respective users of each system) and the bandwidth they offer. On top of the list in the table is the cheapest wireless connectivity technology, the infrared. Infrared ports can provide a very reliable and cost-effective short connectivity (in the range of a few inches to a few feet) between computers, handheld terminals, cellular phones, and peripheral devices such as printers and scanners. Without any kind of infrastructure, these devices can make a computer network or a point-to-point connection at a speed of up to 4 Mbps or more. The second listed system is the wireless LAN, defined in IEEE 802.11 standards. With low-cost APs and access cards, the wireless LAN can provide a very high-speed computer network (such as a LAN or an ad hoc network). Starting with its first version, which provided 2-Mbps data rate, the successive versions of

Table 5.1 Mobility coverage and capacity of different wireless networks

Wireless network	Coverage	Data rate
Infrared	Room	19.2 kbps–4 Mbps
IEEE 802.11	100–500 m around each AP	2–11 Mbps
GSM	Cellular network	9.6 kbps
CDPD (for AMPS[a], IS-95, IS-136)	Cellular network	19.2 kbps
DECT, PHS	Cellular network	32 kbps
GPRS (for GSM)	Cellular network	155 kbps
UMTS/IMT-2000[b]	Cellular network	384 kbps–2 Mbps
Iridium LEO[c] satellite	Global	2.4 kbps
Broadband satellites	Global/regional	2 Mbps

[a] AMPS Advanced Mobile Phone Systems
[b] IMT-2000 International Mobile Telecommunications
[c] LEO low earth orbit

the wireless LAN IEEE 802.11b and IEEE 802.11a can achieve a maximum of 11 Mbps and 54 Mbps speed, respectively. The cost of the system, although very low, is still higher than the infrared but on the other hand it can provide larger mobility range up to a few hundred meters.

In the cellular world, second-generation systems, such as Global System for Mobile communications (GSM), Cellular Digital Packet Data (CDPD), Digital Enhanced Cordless Telecommunication (DECT), Personal Handyphone System (PHS), and the packet-switched 2.5 generation system General Packet Radio Service (GPRS), the successor of the GSM, and finally the third-generation (3G) wireless cellular systems such as Universal Mobile Telecommunications System (UMTS) and cdma2000 offer very large mobility range up to a few miles in radius around a single base station with further coverage expansion through their cellular topology and handoffs between cells. Because of the complexity and infrastructure involved in all these cellular networks (more or less), data service cost is much higher than the first two wireless data systems, the wireless LAN and infrared. However, when we look at the range of data rate they offer to their users, they are tens or hundreds of times below a wireless LAN system.

The last two systems listed in the table are still wireless but utilize the satellites as their base stations. In these systems, even higher initial and running costs are involved compared to cellular terrestrial systems and thus we can see much more expensive service cost to data users. Again, the data rate is too low to be considered as a major breakthrough for today's communications even when considering their global mobility coverage.

The wireless data networks listed in the table and their three QoS metrics show no direct relation. For example, we cannot say that if a user pays more, he can achieve higher mobility range and data rate. For some systems among those listed, it is possible to upgrade the mobility range by paying more (for example, from a wireless LAN to GPRS) but you cannot expect to achieve a higher data rate at the same time. On the contrary, you may get a very low-speed service, too slow to be useful for your particular application that was running in the original network.

This simple example illustrates that although all QoS are important, there must be a limit on the number of these metrics that can be provided to a user simultaneously. The current example might show the present-day technologies and one might say that this limitation is because of the current limitation in technology. However, we can find other examples for which there is no such technological limitation, but it is simply not possible to provide all things under one umbrella of QoS at the same time.

The QoS provisioning and guarantee is all about the trade-off between the many sides of the QoS requirements. Providing all these requirements to users at the same time is not only impossible but also unnecessary and inappropriate for any network. References [10–13] provide further readings on QoS in distributed systems.

5.2 QUALITY-OF-SERVICE GUARANTEE IN IP NETWORKS

In this section, we try to formulate some general ideas on how QoS can be guaranteed in IP networks. The section performs this task by providing simple and very basic examples, but the results are applicable to all kinds of data networks, including our targeted wireless IP network.

The current global Internet service is based on the so-called best-effort service [14,15]. This service does not guarantee any thing other than delivering the IP packets within the network. That is, once a packet is generated and left to the Internet for delivery to a destination host, the network does not guarantee any specified delivery time (delay), the speed at which the packet will be forwarded (data rate and throughput), the available bandwidth for delivering the packet, or even that the packet does not get lost during this delivery. The network only guarantees that it will do its best to deliver the user information with all the resources it has, but if the packet is lost or corrupted, the respective entities should try to retransmit the lost packet and recover it. The IP providing this type of service is said to be unreliable. This unreliable protocol, however, can be overlaid on top of any link layer (e.g. Ethernet or asynchronous transfer mode (ATM)) and that is one of the advantages of the IP. The reliability to the network delivery must be then provided by other layer protocols such as TCP at the transport layer, and therefore it becomes possible to use the Internet.

Internet Engineering Task Force (IETF) [16], the main body in standardization of the Internet, and the Internet research community are working on new proposals to provide better services and some type of QoS support in IP networks. This may start first by services called better than best-effort, with the QoS eventually guaranteed in the future. In such services, IP as the network protocol will have to guarantee some QoS metrics such as delay, average, minimum, or peak throughput, bandwidth, or loss probability for delivering Internet datagrams.

Before going to detailed discussions on these activities, in this section we will see how we can basically establish a QoS network and what our restrictions and requirements are for such a service. We will outline four fundamental principles in providing QoS in data and IP networks. These four principles are common in any data communications network.

5.2.1 Packet classification

Assume a simple example in which a single application is running at two source hosts A and B, respectively, directly connected to a single router R_1. The first one is a delay and

bandwidth sensitive application, such as a real-time voice over IP application that requires around 1-Mbps bandwidth and short delays. The second one is a delay and bandwidth insensitive application such as FTP (File Transfer Protocol) that can tolerate the data rate and delay requirements reasonably. Hosts A and B are connected directly to the router R_1 that is connected to a second router R_2. A limited capacity link of 1.5 Mbps connects the two routers R_1 and R_2. All packets have to be routed through this connecting link to the other parts of the network.

The voice over IP application needs its typical bandwidth and delay requirements in order to have sensible communications. On the other hand, the FTP application can take longer time for delivery of packets so that no restrictive delay or bandwidth requirements are assumed here for this application. This simple example also illustrates different QoS requirements for different applications, as outlined in the previous section.

In an ideal situation, a network manager wants to provide 1 Mbps out of the available 1.5 Mbps to the voice application and the leftover capacity of 0.5 Mbps to the FTP application. One way to do this is to classify and mark different packets (voice packets versus FTP packets) at the input port of the router R_1 so that the router can share the total capacity proportionally to the two hosts. By providing such a share policy, we will be able to provide the QoS to the two hosts connected to our example network. Therefore, the first principle in providing the QoS is that we need to differentiate between different types of packets generated from individual applications.

The currently used version of the IP, IP version 4 (IPv4), has a field in its header called *type of service* (TOS) and the next-generation IP, or IP version 6 (IPv6), also includes a Traffic Class (TC) field within its header. (See Chapter 10 for detailed discussions on IPv4 and IPv6 protocols.) The TOS or TC field in the IP header can, for example, be used for this type of classification or for marking the packets generated from different applications. When the router has the appropriate information on distinguishing between different traffic classes and is equipped with a new discrimination policy according to the packet classes, it will be possible to provide different services to different classes of packets (applications). The current IP, however, rarely uses the TOS field as all packets are handled equally so that we have just a best-effort service. This is a service that allows every user to grab the bandwidth as much as possible without any control or priority consideration.

5.2.2 Packet isolation

The classification of the packets is a very good start in providing QoS in IP networks. But what happens if the applications misbehave and use more network resources than what they really need. For example, if the host A uses more than 1 Mbps for its voice application, it may improve its service slightly but on the other hand it will destroy the host B FTP application performance. So there should be some entity within the network to monitor the behavior of applications and their use of the network resources. Therefore, our second principle here will be that in addition to packet classification, we need to monitor and control that no one uses more resources than what they have been allocated.

The most directly connected router to the hosts (in this example router R_1) can perform this monitoring role. By this method, we will make sure that the task of control and monitoring is distributed within the network and it is assigned to the most appropriate entity in the network. It also provides other benefits usually known for distributed systems (as opposed to central control) such as reliability and minimal exchange of control messaging.

5.2.3 Efficient resource management

One way to provide monitoring that no application uses network resources excessively is to partition these resources. In our example, the bandwidth is the main resource and we can partition it into two parts of 1 Mbps and 0.5 Mbps allocated to host A and host B, respectively. This can be easily done by maintaining two different queues at the input port of router R_1, which can be easily implemented by the software. Note that when we say allocating resources to a user, it does not mean that this allocation is permanent. The allocations can be managed and reconfigured dynamically, for example, only for the duration of the respective application.

One problem that could be raised here is what happens if host A's voice application does not use its allocated resources for some time. This means that although we have tried to discriminate the two applications to provide a good QoS, we may waste the precious network resources. So we come to our third principle that although we need to differentiate individual application's packets and monitor their limit on usage of the available resources in order to provide QoS in our network, we still need to make sure that our resources are not wasted at any time. Therefore, the resource management will be an important issue in any QoS network and we need to use the network resources as efficiently as possible. Management of the queues maintained at the routers could be thought of as part of the overall resource management in the system.

5.2.4 Traffic load control

Our three principles so far are very important and can provide a good service quality in our network. They are also very general and can be used in any network. But we have not included one important factor in our discussions. The fact is that our network resources are limited. In our above example, we had 1.5-Mbps capacity and thus we could share it between two applications, at 1 Mbps and 0.5 Mbps, so that we could provide a good QoS to both the applications. But this is not always the case. Especially, when your network has many users, sharing the available capacity will not be an easy task. For instance, consider our simple example when hosts A and B are both running voice applications and both need 1-Mbps capacity in order to comply with their QoS requirements. Using the simple calculation of 'one plus one equals two', we can see that it is not feasible to share the available 1.5 Mbps between the two hosts and still claim that we provide the QoS as desired by the users.

In such a situation, we can simply forget the QoS and give each user half the available capacity. But if we want to provide QoS, then we need to do something else. That is, we need to allocate to one of the hosts the required 1 Mbps and ask the second one to wait until the required bandwidth is available. The issue of who gets the capacity first and who gets it next is another issue. We may allocate the capacity on a first-in-first-served basis, or on a user subscription priority basis, or any other policy.

Thus, the fourth principle in providing the QoS in a data network is that we need to have Call Admission Control (CAC) in our network (e.g. implemented at routers) in order to handle the situations in which requests for allocation of the network capacity, such as bandwidth, are more than available resources. CAC thus will be a major foundation in any QoS network. The issue of admission control has been widely addressed in many networks for many years (e.g. see References [17–27]) and still there is a lot of work to be done. We will discuss the issue of traffic management, which includes CAC, in more detail in Chapter 7.

5.2.5 Summary

Let us summarize all the principles for providing QoS in a network that we have outlined in this section. As a basic understanding, we should now know that QoS is a term that has to be associated with a particular application run by a user at a given time. The QoS requirements for each application thus might be different from those that are necessary for another application. Therefore, a user needs to have a QoS profile listing appropriate requirements associated with particular applications enabled in the user terminal.

Considering the limited available network resources such as bandwidth, therefore, if we want to provide QoS to a particular user application sharing these resources with other users and applications, we first need to find a way to classify the nature of packets generated by the users. We can mark the packets and show this marking by using some methods such as the field within the IP datagram header to the control entities such as routers.

We then need to monitor the usage of network resources by the users, both to make sure that no one takes more resources than what is assigned to them and also to make sure that these resources are used efficiently.

Finally, we need to make sure that we do not commit some QoS that is more than what we have in hand. So a second control called *traffic control* or CAC is necessary. The admission control policy makes sure that we will accept a new user with some predefined QoS metrics only when we can support these requirements on the basis of the available network resources.

The QoS principles outlined in this section can be considered as the fundamentals of the QoS provisioning in any data packet network. We will use them throughout this book and try to evaluate the QoS provisioning approaches taken by the research communities in different networks to see if they are promising permanent solutions or they are just short-term answers with no future.

5.3 INTERNET SOLUTIONS TO QUALITY-OF-SERVICE PROVISIONING

IETF [16] has started working on QoS in IP networks in the mid-1990s. Two different approaches have been introduced: Integrated Services (IntServ) in 1994 [28] and Differentiated Services (DiffServ) in 1998 [29–32]. Since these services are discussed in many traditional data networks and IP literatures, here we will just review their characteristics in brief for the sake of keeping consistency in our discussions. Readers who are interested in more details of these two services are referred to the references provided at the end of this chapter.

5.3.1 Integrated services

This service has been introduced in IP networks in order to provide guaranteed and controlled services in addition to the already available best-effort service. IntServ is an extension to the Internet architecture to support both non real-time and real-time applications over IP. Each traffic flow in this service can be classified under one of the three service classes:

- Guaranteed-service class
- Controlled-load service class
- Best-effort service

Guaranteed-service class provides for delay-bound service agreements such as voice and other real-time applications, which require severe delay constraints. Controlled-load service class on the other hand provides for a form of statistical delay service agreement, for example, with a nominal mean delay. Finally, best-effort services have been included to match the current IP service mainly for interactive burst traffic (e.g. web), interactive bulk traffic (e.g. FTP), and background or asynchronous traffic (e.g. e-mail).

Guaranteed and controlled-load services are based on quantitative service requirements and require signaling and admission control in network nodes (similar to what we discussed in the previous section). Usually, for these type of services a resource reservation protocol (RSVP) is used [33,34]. RSVP is a signaling protocol used to reserve resources in the routers, in a hop-by-hop basis, considering the applications requirements (e.g. throughput guarantees, end-to-end delay bounds) for a given IP flow.

The main advantages of the IntServ are that it provides service classes that closely match different application requirements; it leaves the existing best-effort service almost unchanged, so that no change will be necessary to the existing applications and they can continue to enjoy the current IP service, and finally, it leaves the forwarding mechanism in the network unchanged, so that nonupgraded networks can still receive data from an IntServ network, without any problem.

On the negative side, the architecture of IntServ requires that for an end-to-end service guarantee, all intermediate nodes need to support the service agreement for a given

Internet flow. Therefore, if in a network somewhere in the middle between the source and destination the service guarantee is not available, then the whole issue of end-to-end guarantee will be lost. In addition, subdivision of the best-effort service may cause problems in commercial networks.

The IntServ and RSVP proposals have failed to become an actual end-to-end QoS solution, mostly because of the scaling problems in large networks and because of the need to implement RSVP in all the network elements from the source to the destination. Because of these scalability problems and other disadvantages discussed above, IntServ will probably never be deployed beyond access networks.

5.3.2 Differentiated services

DiffServ came to remedy the disadvantages of IntServ in providing QoS in IP networks. DiffServ aims at providing simple, scalable, and flexible service differentiation using a hierarchical model. That is, the resource management now divides into two domains:

- Interdomain resource management
- Intradomain resource management

DiffServ also allows the network provider to differentiate different traffic streams, using different per-hop-behaviors (PHB) when forwarding the IP packets of each stream. The advantage of such a scheme is its scalability, since many IP flows can be aggregated in the same traffic stream or behavior aggregate (BA). The PHB applies to an aggregate and is characterized by a DiffServ code point (DSCP) marked in the header of each IP packet. In IPv4 header the TOS field and in IPv6 header the TC field can be used for this purpose, which in DiffServ terms are renamed as DS field, shown in Figure 5.1. PHBs are implemented on IP routers through the management of network resources, namely, of classifiers, markers, meters, queues, droppers, and schedulers. These network resources are managed and allocated to traffic streams according to the provisioning policies of the network provider.

At the local network (now named *local DS domain*), three types of routers can be distinguished: access routers, interior DS routers, and border DS routers. These are shown in Figure 5.2. Access routers are the routers close to the end-user hosts. Several access routers are connected to an interior router. So the interior routers are at the second level far from the end hosts. At the highest level of the local DS domain, all interior routers are connected to a border DS router that connects the local DS domain to the outside world

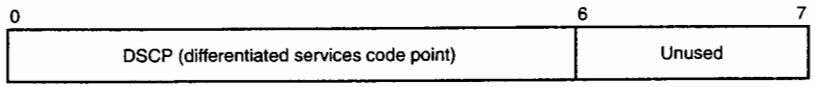

Figure 5.1 DS field in the IP header for the purpose of DiffServ

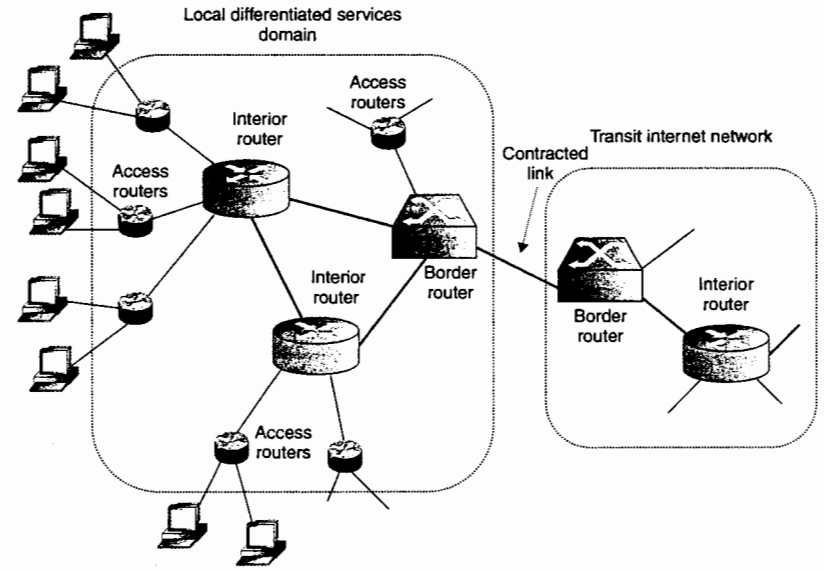

Figure 5.2 Differentiated services network architecture and the three types of DS routers

or the service provider network. The link between the border DS router and the service provider network will be contracted at the DS aggregate rate.

At the interdomain resource management, unidirectional service levels are agreed at each boundary point between a customer and a provider for traffic entering the provider network. At the intradomain resource management on the other hand, the provider is solely responsible for configuration and provisioning of resources within its domain. Therefore, different from IntServ in which all control and resource management have been performed on an end-to-end basis, in DiffServ the local network has to share the resources allocated by the outside network (or service provider) to its users. Scalability, simplicity, and flexibility of DiffServ compared with IntServ come from this hierarchical management.

DiffServ does not impose either the number of traffic classes or their characteristics on a service provider. The provider builds its offered services with a combination of traffic classes, traffic conditioning, and billing.

So, in DiffServ architecture a service-level agreement (SLA) is provided to govern the traffic handling between a local network and the service provider network [35]. After that, it will be the local network that provides the required services to its end users. Per-flow state is also avoided in DiffServ within the network since individual flows are aggregated in classes and will be supported by the local network resource management using the available resources provided on the basis of the SLA.

In summary, in DiffServ the entire customer's local network requirements for QoS are aggregated and then an SLA will be made with the network service provider. The SLA

could be static, that is, negotiated and agreed on a long-term basis (e.g. monthly) or could be dynamic, which changes more frequently. The local network is then responsible for providing DiffServ to end users within the network. This is usually done by marking packets with specific flags shown in the TOS field of the IPv4 or the TC field of the IPv6. For DiffServ purpose, these fields are now renamed DS field and it will supersede the existing definition of IPv4 TOS and the IPv6 TC field.

The provisioning of a DiffServ network is the key for the network to exhibit the expected behavior. Moreover, admission control mechanisms are required at the edge of DiffServ domain in order to avoid network congestion and to prevent QoS degradation. And when this situation occurs, the dynamic reprovisioning of the DiffServ network is required.

Typically, this technology is intended for deployment at the core network, although end-to-end DiffServ is conceivable when the end-to-end chain fully relies on IP DiffServ networks.

5.3.3 Comparison between IntServ and DiffServ

DiffServ comes with some advantages and disadvantages. On the positive side, DiffServ provides the kind of discrimination based on payment for service. Traffic classes are accessible without additional signaling as a traffic class is a predefined aggregate of traffics. Network management will be simpler in DiffServ compared to IntServ, since classification of the traffic needs to be performed at the end systems.

On the other side, DiffServ tries to keep the operating mode of the network simple by pushing as much complexity as possible onto the network provisioning and configuration. DiffServ also does not make the provision of several services with different qualities within the same network easier.

So in summary, IntServ requires flow-specific state for each flow at the routers. State information will be increased in accordance with the number of flows. This will need huge storage space and processing power at the routers and makes routers much more complex.

DiffServ on the other hand is simpler and more scalable compared to IntServ. The reason for the scalability of DiffServ is that the per-flow service is now replaced with per-aggregate service. The complex processing is also now moved from the core of the network to the edge.

5.3.4 IntServ over DiffServ

Recently, it has been proposed to apply the IntServ end-to-end model across a network containing one or more DiffServ regions [31]. Such a proposal fully endorses the use of RSVP as an end-to-end solution to provide QoS in the context of IntServ. Compared to a pure IntServ solution, this approach has some advantages because it removes the per-flow processing from the core routers. However, the per-flow processing remains essential at both the edge and border routers.

5.4 CELLULAR NETWORK SOLUTIONS TO QUALITY-OF-SERVICE PROVISIONING

In this section, we will check the status of QoS establishment in two cellular technologies that are considered as the gateway technologies to the future wide-area wireless Internet. We will see the QoS examples for the 2.5G GPRS and the 3G UMTS systems. This discussion has a twofold purpose. On one side, we would like to see what the cellular researchers think about establishing the service quality to Internet applications that they offer in their systems. On the other side, which is more important than the former, we want to see how the cellular QoS provisioning approach is close (or far) from its Internet counterpart. Future wireless Internet will be based on both technologies, cellular and Internet, and it is very reasonable to think that we need harmony in any improvement attempt toward the wireless Internet from the two major technologies. Without such harmony and cooperation, it would be difficult to think of any reliable wireless Internet system in the near future.

5.4.1 GPRS quality-of-service support

GPRS is the packet version of its well-treated predecessor second-generation cellular system GSM. GPRS is intended to provide a data friendly core and access network that can accommodate data services, mainly Internet types, to the users of cellular networks. GPRS was a result of increasing demand in Internet applications for mobile users and thus it can be considered as the first initiative in wide-area wireless Internet. The advanced features of the GSM network and the new network architecture of the GPRS made it possible that the core network of the 3G system UMTS follows almost a similar concept as its two predecessors.

There is, however, an important fact that we need to note here for GPRS and UMTS. The fact is that both systems are mainly developed by cellular engineers mostly from the International Telecommunication Union (ITU) community with little cooperation from either the Internet community or the IETF.

Implementation of Internet applications, such as e-mail and web browsing, and the increasing demand in providing service quality in telecommunication networks have resulted in consideration of QoS in the new cellular systems.

GPRS defines the QoS requirements for each subscribed user in a QoS profile, which is defined and maintained at the GPRS network Home Location Register (HLR). Serving GPRS Support Node (SGSN) is responsible for fulfilling the user QoS profile at all times, including periods when the user is located outside his home network. (See Chapter 3 for more details of the GPRS network architecture.)

Every subscriber to the GPRS network is allocated a QoS profile that consists of a number of the following QoS indicators:

- Traffic precedence class: defines the priority of service within the network
  - — High, normal, or low priority

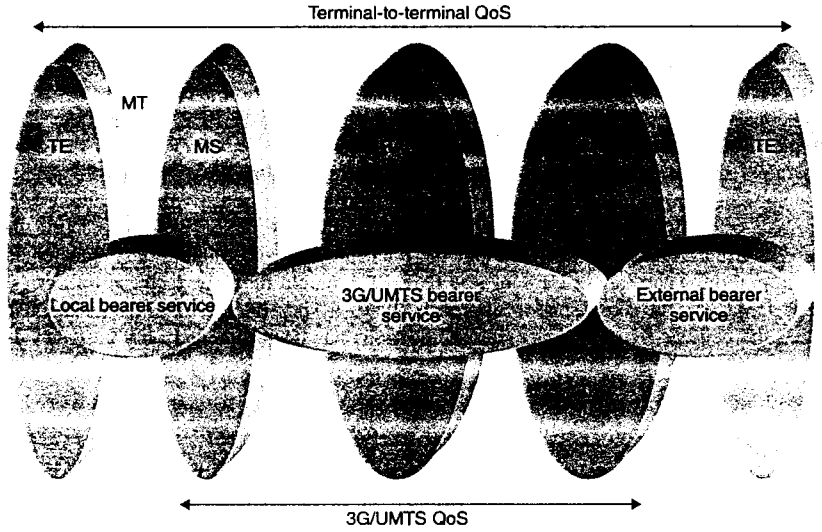

Figure 5.3 End-to-end QoS realization in cellular networks

The discussions given in the previous sections show that, unfortunately, at this time we do not see a close relationship between the Internet and cellular approaches on providing QoS. They have aimed at improving the QoS by their own approaches without paying much attention to the other. Providing end-to-end QoS with such a configuration would be a very difficult task, if not impossible, and therefore it will be a long time before we can see an end-to-end QoS support for the wireless Internet. Harmonization between the two approaches in a way that the service of one technology can cooperate and complement the service of the other remains the main issue toward QoS establishment for future networks.

The simple example above also illustrates that any promise from one system provider in supporting the QoS to its users has no basis as long as the QoS is not supported by all the system providers that complete the end-to-end line. The QoS definitions provided in this chapter should by now have cleared the differences between an end-to-end QoS system and a system with partial QoS support. The latter cannot be considered as QoS support at all.

5.6 SUMMARY AND CONCLUSIONS

In this chapter, we have developed a fundamental understanding of the QoS in telecommunication networks. After defining the meaning of quality, we have concluded that the QoS could not be set as a series of parameters to be used for all networks and all applications.

The QoS should be determined as a subset of required parameters for each network and each application. Therefore, the QoS is application- and network-oriented.

Further, we have developed a set of principles for providing QoS in a communication network, irrespective of the type of the network being mobile or fixed, data or voice. These principles are the requirement of packet classifications, packet isolation, efficient bandwidth utilization, and admission control. These principles can be implemented in any communication network that seeks QoS guarantee.

We have overviewed the Internet and cellular network initiatives for the QoS provisioning. The Internet main directions are the IntServ and DiffServ, with more promising views of the latter, in the context of being practical and feasible. Reservation protocols are named as the means for providing this Internet QoS so that the current best-effort service can be evolved into a guaranteed service.

For cellular networks, the QoS provisioning is discussed in different ways in the literature. There is little in common between the cellular QoS initiatives and the Internet approaches. Examples of GPRS and UMTS are given for the cellular part.

It is concluded that the diverse approach taken for providing the QoS in the Internet and in the cellular mobile communication systems could cause major problems in future wireless IP networks. The wireless IP needs both technologies as its foundation and taking different approaches on the same issue in the network would never provide an efficient resolution. Moreover, this will make the end-to-end QoS guarantee very difficult and costly, and without an end-to-end QoS solution, we cannot claim any victory in achieving QoS.

We will address QoS on many occasions in the following chapters in this book. Chapters on traffic management (Chapter 7) and mobility in cellular network (Chapter 8) are closely related to this important issue but other chapters will also go through the QoS provisioning in relation to the subjects in perspective (e.g. ad hoc networks in Chapter 12 and satellites in Chapter 13). Therefore, it is simply not possible to limit discussions on QoS to what has been provided in this chapter, and neither will the researchers in the field stop doing research on this subject, and, therefore, the research on QoS will have a long way to go.

REFERENCES

1. Chalmers D & Sloman M, A survey of quality of service in mobile computing environments, *IEEE Communications Surveys*, Second Quarter, **2**(2), 2–10, 1999.
2. Guerin R & Peris V, Quality-of-service in packet networks: basic mechanisms and directions, *Computer Networks* **31**, 169–189, 1999.
3. Aurrecoechea C, Campbell AT & Hauw L, A survey of QoS architectures, *ACM Multimedia System Journal*, Special Issue on QoS Architecture, May, 1998.
4. Knoche H & de Meer H, Quantitative QoS-mapping: a unifying approach, *Proceedings of 5th IFIP International Workshop on QoS*, 1997.
5. Hutchison D, Mauthe A & Yeadon N, Quality of service architecture: monitoring and control of multimedia communications, *Electronics and Communications Engineering Journal*, **9** (3), 100–106, 1997.

6. Das SK, Sen S K, Agrawal P & Basu K, Modeling QoS degradation in multimedia wireless networks, *Proc. IEEE Int. Conference on Personal Wireless Communications (ICPWC)*, Mumbai, India, December 1997.
7. Das SK & Sen SK, Quality-of-service degradation strategies in multimedia wireless networks, *Proc. IEEE 48th Vehicular Technology Conference (VTC '98)*, Ottawa, Canada, May 1998, pp. 1884–1888.
8. Nahrstedt K & Steinmetz R, Resource management in networked multimedia systems, *IEEE Computer*, **28**(5), 1995.
9. Campbell A & Coulson G, A QoS adaptive multimedia transport system: design, implementation and experiences, *Media Distributed Systems Engineering*, Vol. 4, 1997, pp. 48–58.
10. Hutchison D et al., In Sloman M (Ed.), *QoS Management in Distributed Systems in Network and Distributed Systems Management*, Addison-Wesley, Reading, Mass., 1994, pp. 273–302.
11. G Bochmann, and A Hafid, Some principles for quality of service management, *Distributed Systems Engineering*, Vol. 4, IOP Publishing, 1997, pp. 16–27.
12. Das SK, Jayaram R & Sen SK, An optimistic quality-of-service provisioning scheme for cellular networks, *Proc. IEEE Int. Conference on Distributed Computing Systems (ICDCS)*, Baltimore, Md., May 1997, pp. 536–542.
13. Blair G & Stefani J-B, *Open Distributed Processing and Multimedia*, Addison-Wesley, Reading, Mass., 1997.
14. Xiao X & Ni LM, Internet QoS: a big picture, *IEEE Network*, March/April, 8–18, 1999.
15. Fry M et al., QoS management in a world wide web environment which supports continuous media, *Distributed Systems Engineering*, Vol. 4, 1997, pp. 38–47.
16. Internet Engineering Task Force, Web site: http://www.ietf.org.
17. Knightly EW & Shroff NB, Admission control for statistical QoS: theory and practice, *IEEE Network*, March/April, 20–29, 1999.
18. Kim J & Jamalipour A, Measurement-based admission control for wireless IP networks, *2002 International Symposium on Performance Evaluation of Computer and Telecommunication Systems SPECTS 2002*, San Diego, Calif., 14–19 July 2002, pp. 587–591.
19. Grossglauser M & Tse D, A framework for robust measurement-based admission control, *Proc. ACM SIGCOMM '97*, Cannes, France, September 1997.
20. Gibbens R, Kelly F & Key P, A decision-theoretic approach to call admission control in ATM networks, *IEEE Journal on Selected Areas in Communications*, **13**(6), 1101–1113, 1995.
21. Jamin S, Shenker S, Danzig P, Comparison of measurement-based admission control algorithm for controlled-load service, *Proceeding of IEEE ICCS 1997*, Kobe, Japan, April 1997, pp. 973–980.
22. Jamin S et al., A measurement-based admission control algorithm for integrated service packet networks, *IEEE/ACM Transactions on Networking*, December 1996.
23. Cheng L, QoS-based on both call admission and cell scheduling, *Computer Networks and ISDN Systems* **29**, 555–567, 1997.

24. Knightly E & Shroff N, Admission control for statistical QoS: theory and practice, *IEEE Network*, March/April, 20–29, 1999.
25. Misic J, Chanson S & Lai F, Admission control for wireless multimedia networks with hard call level quality of service bounds, *Elsevier Computer Networks*, **31**, 125–140, 1999.
26. Peha J, Sched0.uling and admission control for integrated services networks: the priority token bank, *Elsevier Computer Networks*, **31**, 2559–2576, 1999.
27. Ayyagari D & Ephremides A, Admission control with priorities: approaches for multirate wireless systems, *Mobile Networks and Applications*, **4**, 209–218, 1999.
28. Braden R, Clark D & Shenker S, *Integrated Services in the Internet Architecture: An Overview*, RFC 1633, June 1994.
29. IETF DiffServ Working Group Charter, RFCs and Internet Drafts: http://www.ietf.org/html.charters/diffserv-charter.htm.
30. Blake S, Black D, Carlson M, Davies E, Wang Z & Weirs W, *An Architecture for Differentiated Services*, RFC 2475, December 1998.
31. Bernet Y, Ford P, Yavather R & Baker F, *A Framework for Integrated Services Operation Over DiffServ Networks*, RFC 2998, November 2000.
32. Nickols K, Jacobson V & Zhang L, *A Two-Bit Differentiated Services Architecture for the Internet*, RFC 2638, July 1999.
33. Braden R, Zhang L, Berson S, Herzog S & Jamin S, *Resource Reservation Protocol (RSVP) - Version 1-Functional Specification*, RFC 2205, September 1997.
34. Zhang L et al., RSVP: a new resource reservation protocol, *IEEE Network*, **7**(5), 1993.
35. *Service Level Specification Semantics and Parameters*, IETF Internet Draft (work in progress), February 2002, http://search.ietf.org/internet-drafts/draft-tequila-sls-02.txt.