CHAPTER 4

General Framework for Differentiated Services

This book tries to give as consistent a view as possible on the Differentiated Services effort—its targets, available methods, and the most promising service models. One way to meet this objective is to build a framework based on an evaluation of various aspects. The framework built in this chapter attempts to have the same meaning for Differentiated Services as grammar has for a language: "Essence is expressed by grammar," Ludwig Wittgenstein once remarked (Wittgenstein 1973).

Although the introduction section of this chapter is quite extensive, it is important to remember that all considerations are inevitably based on simplified models of complex reality, and some aspects are inevitably ignored. The real challenge of this whole effort is to acknowledge the many relevant issues that could be considered, while realizing that there is no hope of a grand unified theory that can explain everything. Even taking into account this inescapable limitation, however, the ensuing evaluation may provide useful tools for understanding and building Differentiated Services.

This chapter first attempts to clarify the philosophical standpoint of this book by analyzing the complexity of the matter and the most fundamental concepts of service differentiation. Then the chapter briefly reviews the basic limitations of the documents made within the Differentiated Services Working Group. Based on this evaluation, a target is set for a more general framework.

A closer examination is made of three issues: the efficiency of statistical multiplexing, the need for high predictability of quality, and a tool that can be used to compare service models from guaranteed services to relative service. Finally, a concise, but general framework with three primary dimensions is introduced.

4.1 Basis for the Framework

This section defines the position of the Differentiated Services framework introduced in this chapter. Together, three parts make up this framework:

- The outer limit defined by the fundamental concepts
- The inner limit defined by the scope of the Differentiated Services Working Group
- The general target for the framework

Two fundamental terms, *differentiation* and *service*, together define the extreme limits for this effort. All the issues presented in this book and the framework will be related to service differentiation. It is so vast an area, however, that it cannot be exhaustively covered.

The scope of the Differentiated Services Working Group defines the minimum region for the framework. Nevertheless, that region is quite limited because of some fundamental restrictions of a standardization organization. Particularly, service and business models are largely excluded from the specifications made by the Working Group.

The reasonable area for the framework is situated somewhere between the outer and inner limits. The primary purpose of the framework is to facilitate the designing and implementation of Differentiated Services. Because the ultimate goal is to provide service for customers, however, service and business aspects should be essential parts of the framework.

It also is useful to remember the large number of issues discussed in Chapter 1, "The Target of Differentiated Services." A pivotal requirement for the framework is that it should be applicable to various cases in a consistent manner.

4.1.1 Basic Terminology

It is important to first consider the meaning of the primary term of the whole book: *differentiation*. One definition for differentiation can be found in the glossary of *The Origin of Species* (Darwin [1859] 1972): "the separation or discrimination of parts or organs that in simpler forms of life are more or less united." This definition is surprisingly usable in the context of the Internet, where differentiation may mean that the current, simpler form of service is separated or discriminated into several forms. The Oxford Dictionary, which offers a more recent definition of differentiate, says it is "to constitute the state of being distinguishable in nature, form, or quality." These three attributes provide a good starting point also for considering service differentiation.

The meaning of *service* may seem evident. However, one general remark is useful: Service is not necessarily something that is truly sold, nor is any specific mechanism necessary to limit the use of the service. Indeed, one characteristic of the original Internet service used

by the academic community was that it was available free of charge for the end users. Users were, however, supposed to carry out research that benefited the community. It is useful to remember this background because it has significantly affected the way the Internet is still working.

When the word *differentiate* is added, some changes in the service model are inevitable. First, there seems to be a strong incentive to attach different prices to different services (where price could also be something other than money). Therefore, the innocent-looking word *differentiate* may yield fundamental changes in Internet service.

4.1.2 Limitations of the Differentiated Services Working Group

The Differentiated Services Working Group made noteworthy progress during the first year after its establishment, in particular taking into account the complexity of the questions and the different views about the target of the effort. However, the results do not cover all relevant issues in the form they are presented. You can find these results in RFC 2474, "Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers" (Baker *et al.* 1998), RFC 2475, "An Architecture for Differentiated Services" (Black *et al.* 1998), and some Internet drafts. There seem to be three primary reasons why the results aren't all inclusive:

- Some issues cannot be addressed formally in an IETF working group (business models).
- Issues are left open mainly because it is better to obtain more experience before making any final decisions (service models).
- There is a lack of common understanding about some fundamental issues, such as the real need for per-flow guarantees.

Customer service paradigms and pricing models are often kept out of the standardization process, because an official document that gives too much guidance relating to these business aspects is seen as potentially limiting competition in the area. Fortunately, because this book is definitely not a standard, it is possible here to address to some extent the delicate area of business models for the Internet. It should be stressed, however, that all suggestions and recommendations presented in this book are based purely on this author's best understanding of the issue, and they are not aimed to limit the application of any kind of Internet business model. Nevertheless, without any business considerations, the usefulness of this book might be seriously deteriorated.

It is hard to decide exactly when the time is right for making a standard. In the beginning, there is not enough understanding about the target and the best mechanisms to meet the target. Then, all of a sudden, it might be too late to devise a standard because someone

has already brought a successful product to market without any standardization. One approach for solving this dilemma is to first make a quite loose standard that defines the frame for products or services without specifying the details of implementations. Later, when more experience is gained, it is possible to refine the standard. Many IETF working groups employ this basic approach (including the Differentiated Services Working Group).

The Differentiated Services Working Group, as explained in Chapter 3, "Differentiated Services Working Group," deliberately leaves many technical issues for further study or to be decided by service providers and network operators. The Working Group's reasoning is that this approach encourages the development of a wide variety of Differentiated Services without limiting the scope of acceptable models. Although this is certainly a reasonable approach, and this book also emphasizes the need for experiments and trials to evaluate different models, there seems to a need to give more comprehensive guidance for implementation even before thorough practical experience (even trials should be based on some understanding).

Finally, even though there seems to be relatively wide consensus that something like Differentiated Services could be very useful for the future Internet, there is still much controversy about how it should be done and what is the right basis for development. The main effect of the growing consensus seems to be that new experts are ceaselessly coming to the field of Differentiated Services. Every new person has his or her own opinion about how things should be done—although this is quite natural, it yields an acute problem of making any progress difficult because the discussion circles the same topics time after time.

4.1.3 Target of the Framework

In short, the target of the framework presented in this chapter is to provide better order in the wild field of Differentiated Services. The ordering is not accomplished by punishing for wrong acts or opinions, but rather by harmonization of comprehension. Harmonization can, however, be seen as a double-edged sword: In addition to the benefits of harmonized systems, there could be the danger of limiting the variability of ideas. To avoid this pitfall, the framework should be, in addition to mandatory consistence, as versatile as possible.

The task at hand can be divided into three phases: to lay the foundation, to erect the skeleton of the building, and to add all other necessary elements.

First, a solid grounding is needed. Because of the lack of common understanding of some fundamental issues, a lot of effort is made in this chapter to build a solid and legitimate basis. This phase focuses on issues such as the efficiency of statistical multiplexing and the usefulness of network service for different applications.

Second, a comprehensible and consistent structure is required. For this, it is important to identify all the integral building blocks and the main relationships among them. As for

Differentiated Services, this means that the purpose of the six bits in the DS field must be clearly defined. Moreover, it is not enough to define the meaning of a single bit combination, or codepoint, but it is of great importance to define the structure of relationships among codepoints. Finally, the outcome is practicable in the sense that it is applicable to solving various practical problems. Based on the foundation and skeleton, it must be possible to construct Differentiated Services (and that should be more useful than a grand monument similar to some former network services).

Recall the attributes introduced in the first chapter of this book: cost efficiency, robustness, versatility, and fairness.

As discussed in Chapter 2, "Traffic Management Before Differentiated Services," neither a pure best-effort model nor a pure guaranteed-service model can provide an efficient solution in a multiple-service environment. In fact, a combination of high-quality requirement of some flows and a highly variable traffic process of some other flows tends to result in low utilization, if there is only one service class. If a large number of service classes are used, the management overhead tends to increase and impair cost efficiency. But how should these considerations be reflected in the framework of Differentiated Services? A clear and consistent structure and avoidance of any useless mechanisms seem to be the key means to reach this target. The same conclusion is largely valid with robustness: A logical and solid structure is the tool that enables the building of a robust system.

Because the main purpose of the framework is technical—that is, the design and implementation of Differentiated Services—it is reasonable to emphasize technical aspects. Because the ultimate goal is to provide service for customers, however, consideration should also be given to marketing and customer care. In the best case, the framework could be useful for both marketing and supporting the service. That is possible, however, only if the framework is really clear, almost self-evident. One should check, from time to time, whether it's possible to convince ordinary customers of the fairness of certain characteristics of the service structure. Or more concretely, the service providers should be able to explain to their customers why they should pay more for one service than for another service.

If IP and the Internet dominate the future of communications, as it now seems, all imaginable, and even not yet invented, services have to be supported by IP. The view presented in this book supports the idea that this versatility requirement should be met by one all-encompassing service model rather than by a collection of separate services—and the most promising versatile service model is Differentiated Services.

4.2 Tools for Evaluating Service Models

This chapter elaborates some fundamental issues of service provision and service differentiation. The general target is to offer as wide a perspective as possible in the sense that substantially different service models can be evaluated with the same concepts and within the same framework. To reach this target, questions such as these should be answered:

What does better service mean?

Which fairness aspects are relevant in case of several layers of aggregation?

A primary concept for accomplishing this task is "availability of quality." The following pages introduce the concept and give some numeric examples to illustrate its applicability. Moreover, another aspect of fairness is assessed—that aspect related to multiple levels of differentiation (between parts of the application, between users, and so on) as presented at the beginning of this chapter.

4.2.1 Availability of Quality

One of the integral issues related to service differentiation is the actual meaning of "better service." Better service seems to be used as a synonym for higher quality and, surely, they have a close relationship with each other. There is an apparent danger of making the wrong conclusions, however, if these terms are taken as exact synonyms. The term *higher quality* may readily lead our thinking to an analysis of a situation in which a connection through the network is already available, and only the quality of the connection is assessed. This kind of viewpoint could be feasible if there were significant pricing only when the service is really used by the customer. On the contrary, in case of flat-rate pricing without a direct relationship between actual use and price, this kind of approach may lead to a peculiar conclusion: Quality can be improved just by arbitrarily rejecting service totally to some customers. Therefore, better service is the broader target than high quality.

Instead of merely using the term *quality*, this book applies a concept of the *availability of quality*. In practice, this term is used in such a way that quality is defined in some traditional terms, such as packet-loss ratio, available bandwidth, and maximum delay; then the availability of that given quality is calculated or measured. Thus availability of quality is the probability that the network service can meet a given quality requirement.

The basic idea is very simple: Service is available for an application only if certain quality criteria are fulfilled; otherwise, the service is unavailable. For a data application requiring an attainable bit rate of at least 50kbps over 10 seconds, availability could be high, say 99.9%, even though the service is only best effort. The availability of the same best-effort service is much lower, perhaps 50%, for an IP telephony application that requires a delay variation of less than 50 milliseconds. Consider the following numeric example.

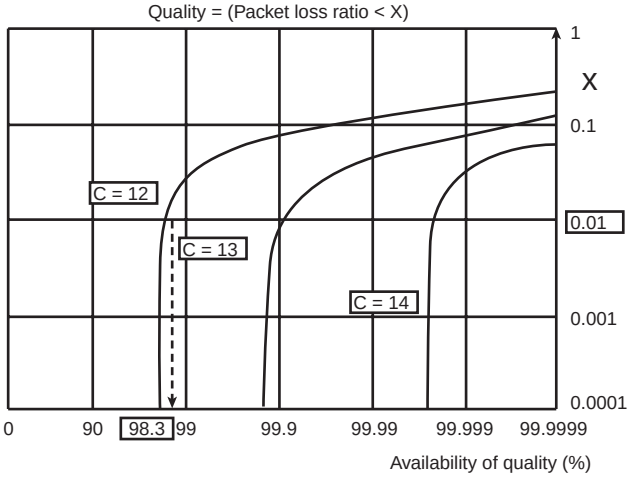
This simple example evaluates the availability of quality on one link used by a number of customers. Based on long experience, the operator knows that the average load is 10Mbps, whereas the network operator does not know beforehand what is the actual momentary bit rate (say, within 10 seconds). Based on the same experience, the uncertainty of the traffic prediction can be described by a variance of $(1\text{Mbps})^2$. In other words, even though the operator knows that the average load is 10Mbps, there are unpredictable variations with variance of $(1\text{Mbps})^2$.

To make some progress, it is necessary to suppose something about the bit-rate distribution. Supposing that the average bit rate measured over 10 seconds is normally distributed, it is relatively easy to calculate the momentary packet-loss ratio for a given link capacity by assuming that all packets within the capacity are transmitted and all excessive packets have to be discarded. Note that this simplified model ignores the effects of short timescale variations and buffering.

Figure 4.1 shows the result in case of three link capacities: 12, 13, and 14Mbps. The vertical axis illustrates the quality criterion, in this case, the packet-loss ratio ($_p$); the horizontal axis is the availability of the quality. With a capacity of 12Mbps, for instance, a packet-loss criterion of 1% yields availability of 98.3%.

Figure 4.1 Availability of quality for a packet-loss ratio.

I updated the fig but dont see any change, Did it go through art and Illustration?



One conclusion seems to be evident: In this case, it is almost irrelevant whether the quality criterion is P~loss = 10^{-4}, 10^{-6}, or even 0, because almost always $_p$ is either 0 or more than 10^{-4}. The quality of service should, therefore, be specified as the availability of certain quality criteria rather that as the average packet-loss ratio. Table 4.1 provides some numeric values that further illustrate the situation. In particular, note the tiny differences between the figures on the first three rows.

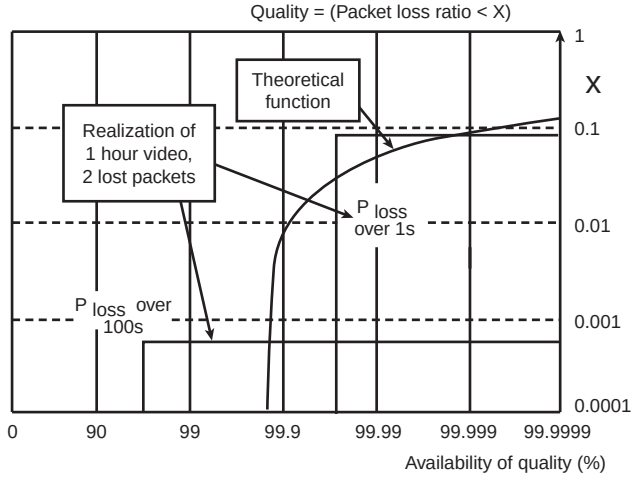
Table 4.1 Availability of Quality for Different Packet Loss Ratios

Quality Criterion	Capacity Mbps C=12	C=13	C=14
$_{P} = 0$	97.720	99.864	99.9968
$_{P} < 10^{-6}$	97.720	99.864	99.9968
$_{P} < 10^{-4}$	97.726	99.865	99.9968
$_{P} < 10^{-2}$	98.300	99.913	99.9983

It should again be stressed that a simple illustrative model cannot cover all relevant aspects. In reality, the dynamics of the packet losses could be very complicated—the human perception of quality can be complicated also. Figure 4.2 illustrates this phenomenon by showing the availability of quality for one fictitious video stream with a length of 1 hour, an average bit rate of 100kbps, and an average packet size of 500 bytes. Supposing that two packets were lost in all, the average packet-loss ratio measured over the total duration is $1.1*10^{-5}$.

If the packet-loss ratio is measured over a period of a second, and the packets are lost during the same second, the packet-loss ratio is either 0 (during 3,599 periods), or 8% (one period). Therefore, the availability of quality is 3599 / 3600 = 99.97% for any packet-loss criterion less than 8%. If the average packet-loss ratio is measured over a period of 100 seconds, the result differs remarkably: Availability is only 35 / 36 (97.2%) for any value of packet-loss ratio less than $8*10^{-4}$, and 100% for any packet-loss ratio greater than $8*10^{-4}$.

Figure 4.2 Availability of quality with a theoretical curve and two practical examples.



Without further information on the dynamics of traffic variations and on the effects of packet loss to the perceivable quality of the applications, it is impossible to say which one of the alternative results is most relevant. Nevertheless, the concept of availability of quality, as described in the following sidebar, can be used to illustrate characteristics of services better than what is possible with average quality parameters.

Criterion for Quality Availability

It is possible to make the evaluation more concrete by assuming that the customers of the fictional service provider, Fairprofit, are using one link for real-time applications that require a packet-loss ratio of 10^{-6}. What is the required link capacity if the average bit rate and bit-rate variations are the same as in the case with M = 10Mbps and V = $(1\text{Mbps})^2$ illustrated in Figure 4.2?

The service provider cannot answer this question without specifying the availability criterion. For instance, availability of 99.99% means approximately 1 second of unavailability in every 3 hours. If the service provider believes that this is good enough for customers, a reservation of 13.7Mbps/s is sufficient. What is gained by changing the packet-loss criterion from 10^{-6} to a very high 10^{-2} while keeping the availability criterion the same? According to the present model, 13.6Mbps is needed, and that is only 100kbps less than originally required.

If the service provider looked at the average packet-loss ratio only, the result would be totally different. A theoretical (long-term average) packet-loss ratio of 10^{-6} is achieved by a capacity of 13.8Mbps, whereas a packet-loss ratio of 1% is achieved by a capacity of 10.8Mbps. That is not necessarily the best method for network dimensioning, however, because the situation is most probably that the packet-loss ratio is either 0 or very high for a short period.

4.2.2 Levels of Aggregation

One of the fundamental obstacles in the application of the Differentiated Services model is that the mechanisms inside the network are related to individual packets or aggregate streams, but the desired behavior is usually best specifiable in terms of flows—that is, a sequence of packets transmitted by one application. In addition, there are other important levels of aggregation: the customer, possibly using several applications at the same time; and the organization, which may pay for the service used by the end users.

The desired effects of all these levels should be mapped into a couple of PHB, and that seems to be almost an impossible mission. One way to make the entire system work is to identify the requirements of different entities, and then compare them in terms of desired characteristics and fairness:

- In a pure application model, applications need enough resources for sufficient quality (but waste of resources is undesirable).
- In a customer model, each customer should get a fair amount of resources relative to the price.

- In an organization model, organizations should get the right amount of total resources relative to the total price and support of optimal use of resources within the resource pool.

If you had to support only one of these basic cases, the task of service differentiation could be relatively easy. The reality is that in some parts of the network all these requirements must be satisfied simultaneously. It might be helpful, however, to start by looking briefly at each individual case.

The perspective of one application is somehow very attractive—what else could be the target of a network than to meet the requirements of each application? A natural approach is to list the requirements of all known applications and then design a network that can satisfy all imaginable quality requirements of all applications. This approach seems to lead to a system with one flow versus n–1 flows, where n is the number of all active flows in the network. A good service means that one flow is somehow protected from the effects of other flows in a way that high enough quality is available for every individual flow. If an application needs more bandwidth or better quality than another application, it is fair that it gets better service, isn't it? Yet, it is somewhat difficult to assess the real fairness because an application as such is unemotional and does not usually make any significant decisions.

Another possible perspective is that the network provides a quite rudimentary service structure without any tight connection between any service model and any application. The customer is allowed to use any network service in any way (for instance, a real-time network service for data transfer or vice versa). In this customer model, the fairness criterion is whether an individual customer gets a fair amount of bandwidth and quality relative to the price she is paying. This fairness criterion seems to be almost opposed to that of the application model.

Which one of the criteria is more important in cases where there is an actual conflict? Apparently, the answer depends on the service model provided by the service provider. Supposing that all customers behave agreeably and do not waste resources, the application model may yield a better overall result. In case of egoistic customers, however, the application model is more vulnerable than the customer model.

In the third case, with a number of end users within an organization that pays for the service, it is possible to apply both the application model and the customer model. The total capacity should be divided among organizations based on the prices (customer model), while the capacity of one organization divided based on requirements of applications. The model shown in Figure 4.3 summarizes this introduction to fairness issues. The fairness aspects in the figure are as follows:

F1: Between two applications used by the same end user

F2: Between two applications used by different end users within the same organization

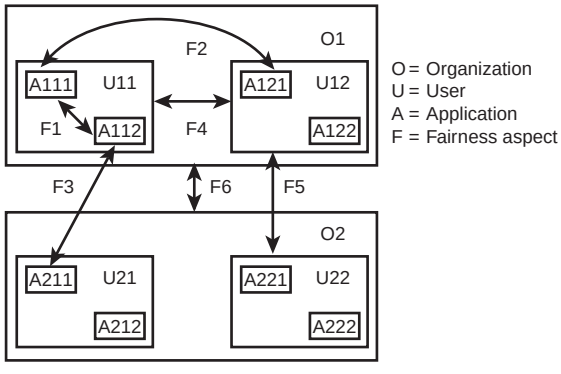
F3: Between two applications used by different end users belonging to different organizations

F4: Between two end users within the same organization

F5: Between two end users belonging to different organizations

F6: Between two organizations

Figure 4.3 Fairness aspects on different levels of aggregates.



Based on this model, it is possible to define three service models according to the significance of the different fairness aspects:

Application model: F1=F2=F3>0, F4=F5=F6=0,

End-user model: F4=F5>0, F1=F2=F3=F6=0,

Organizational model: F1=F2>0, F6>0, F3=F4=F5=0

Fx > 0 means that the fairness aspect is essential for the service model, Fx = Fy means that the two relationships are managed equally, and Fx = 0 means that the aspect is ignored in the service model. In addition, there are different variations of these basic models—for instance, if end users have different rights for using the networks resources in the organizational model, then F4 > 0.

4.3 Customer Service

The first chapter of this book emphasized the fairness aspect of customer service. Although it is definitely a significant aspect, it may not cover the whole field of customer services. A

service could be fair without much attractiveness, or attractive without much fairness—sometimes the service model could be even intentionally vague. In any case, it seems that to be successful there must be some basic attractiveness in the service.

4.3.1 Fulfilling Consumer Expectations

One practical issue to be addressed by service providers is how to build a network service that can fulfill consumer expectations. The following statements are relevant for an ISP providing Differentiated Services:

- Most users are not particularly interested in technical details of network service (that is, bit rates and milliseconds), but in the contents of the application. Those users are either incapable or reluctant to spend time appraising the selections of complex services.
- On the contrary, technologically oriented people may appreciate the detailed service offering with perhaps 100 different bandwidth and quality classes.
- One important aspect for an ordinary end user is the freedom to select the destination (or the source of information flow) without considering anything, such as price or the geographical location of the other end. A service model with these characteristics can be called *unscoped* service.
- On the contrary, big organizations may definitely want to limit the number of destinations, or the scope. The separation of traffic streams improves the possibility to offer virtual private networks with special characteristics, such as security and high quality.

Note

A *virtual private network (VPN)* is a network established for the exclusive use of a single organization with an emphasis on privacy and reliability. The main advantage of a VPN compared to leased lines is that it can exploit the statistical gains and scale advantage provided by large public networks.

These examples illustrate some of the difficulties faced when a service provider wants to fulfill the expectations of all customers. The expectations and preferences could be totally opposite, but still the service provider should use the same infrastructure to provide all kinds of service. A lot of versatility is clearly needed to make the total offering attractive for different type of customers.

4.3.2 Pricing Models and Predictability of Quality

What are the main service characteristics that users are paying for? High bandwidth, low packet-loss ratio, and small transmission delay are the first issues that come to mind, and

certainly they are important issues. Then there is one additional aspect that is not always evident: predictability of quality. The essence of predictability of quality, or lack of it, is that even though the average quality could be high, users might be dissatisfied if they cannot predict the quality level in advance.

You have called your friend abroad using IP telephony service, for instance, and the quality of the connection is excellent during the first minute. Then, suddenly, the quality drops below satisfactory for some seconds and then returns to an excellent level. Although the average quality is probably good by most measures, you may feel uncomfortable because you don't know in advance what will happen in the next minute.

It seems likely that users are willing to pay for a clearly predictable result. To get a preliminary insight of this complex issue, consider two situations: one with time-dependent pricing, and another with flat-rate pricing. Let's try to assess predictability requirements at different points of time.

One Month in Advance

You want to be assured that you can get a certain service at a certain price in the future. You need this assurance to rationally select your service provider. In case of time-dependent pricing, both quality of service and price are predictable: You can get quite reliable information from the Web pages of service providers.

In a flat-rate case, the price is definitely predictable, but service level is not. You need some kind of understanding about the service offered by different providers. You may select the provider just based on price, but more likely you compare service levels of different providers by discussing with your friends or by reading articles about the issue. Still, you want to obtain reliable information that makes it possible to predict what you will really get.

A Few Days in Advance

You make a decision whether to reserve time for something probably one or a couple of days in advance. For instance, you decide to have an interactive meeting with voice and video over the Internet tomorrow.

In case of time-dependent pricing, you know the price, and you likely can be sure that you will get the service. Your decision could be quite straightforward: If you think that the meeting is useful enough compared to the price of the service, you and your colleague allocate time for it.

The flat-rate case seems to be more difficult, because you might not be as sure about the availability of quality. You have some experience, and suppose that things are going in the same way as earlier, and make the decision based on those experiences. The main cost of

this decision is not directly money, but time. Wasted time could actually be very expensive, particularly when it means that several persons have allocated time for the meeting days in advance: It could be extremely irritating if five persons have managed to reserve two hours for a meeting and then it must be cancelled because of a technical reason.

Decision to Buy

In many cases, you make the decision to use a service, such as a telephone call, just before you try it. You want to know, at least approximately, how expensive a service is before you make a decision whether to use the service (or at least to test it). With time-dependent pricing the procedure is quite clear: Because you know both the quality and price, you have the ingredients for a rational decision.

In case of flat-rate pricing, this decision does not seem to be relevant because the actual use of the service is free of charge. The main issue probably is whether you make some kind of reservation for the use of the service. The reservation could be soft in the sense that you just have an idea that you want to discuss with your friend about the program of next weekend. Even in this case, an unsuccessful result would be annoying, although you are not spending any money for trying the service.

Decision to Continue

The final decision of whether to spend time with the service happens some time after you have started to use the service. The worst situation from your perspective is when the quality drops below acceptable in the middle of the service, so that you have spent your time (and possibly money) without a satisfactory result. Predictability is, again, needed for a reasonable decision.

With time-dependent pricing, your anticipation probably is that the quality is constant; so you can quit soon after the start, assess whether the quality is high enough for your purpose, and determine whether it corresponds to the price.

Then with flat-rate pricing, your expectations may be somewhat different. A natural model for service usage is that you test the service first, and then decide whether you want to use it. You may perhaps not need to be sure that the quality is always available, but you definitely want to avoid situations where the quality drops suddenly after a while.

4.3.3 Provisioning High Predictability

This section assesses the possible advantages of having guaranteed services with advanced admission-control mechanisms. Two possible answers are as follows:

- The availability of quality could be better during overload situations with guaranteed services. (This issue is discussed later in section 4.4.4, "Improving Statistical Multiplexing.")
- In certain cases, there is a need to keep the service quality of an individual connection more predictable than what it inherently is without any additional mechanism.

It seems that in the case of pricing of individual flows, it is right and fair that the existing flows get some level of priority over new requests. The primary argument for doing this is the advantage attained by the user of the service: Higher predictability of quality makes the service more useful and more attractive. The problem is that although this reasoning is valid with those applications that definitely require a certain minimum bandwidth, it seems to be questionable with the majority of data applications. Therefore, that kind of priority should not be applied to all flows in the Internet.

In cases where pricing is based on a flat rate for individual users, and where applications can be adjusted easily to a different load condition, it is not at all clear that existing connections deserve any priority over any new connection request. Think, for instance, of a situation in which you cannot get a connection to the Internet because your neighbor has a permanent high-bandwidth connection through the same access link. What could be a justification to give priority for that existing connection?

One interesting question is whether it is possible to provide high predictability in a Differentiated Services network without capacity reservation for individual flows. It seems difficult to do anything inside the network if the boundary node is not giving appropriate information. (Remember that flow states are not kept in core nodes, which makes it impossible to keep track of the duration of each flow.) Basically, this situation leaves two alternatives:

- New connections are accepted only if it is certain that there is enough capacity throughout the network and for the whole duration of the flow.
- Packets belonging to existing flows get higher preference than new flows—or more generally, the preference may depend on the time the flow has been active.

Although it could be possible to implement both approaches using the Differentiated Services system, it is not clear whether either of them is useful in practice.

4.4 Operation and Management

As discussed earlier, a prevalent view seems to be that an ideal QoS provision means that every flow is protected from all external effects in a way that the quality always satisfies the

predefined requirements. The situation from one application viewpoint is 1 versus n–1, where n is the number of all active flows, and n could be millions. From an operator viewpoint, the situation is that of N times (1 versus n–1). That means that the network management can hardly be based on an approach where every individual flow is protected from the effect of every other flow. Something more systematic is needed, and this chapter tries to meet that target.

4.4.1 Predictability of Load and Destination

Let's start with an overview of network services. As stated in the beginning of this chapter, the differentiation may concern nature, form, or quality. What do these aspects mean in the case of Differentiated Services? This issue can be divided into three parts: issues related to destination, traffic, and quality.

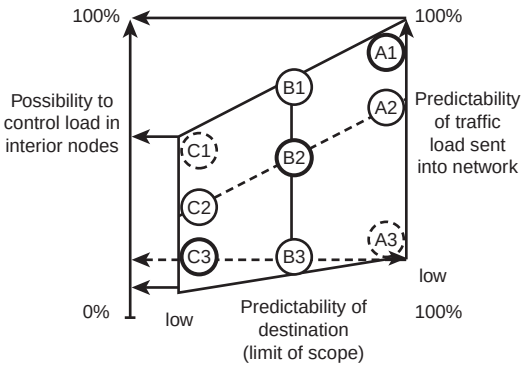
The service may differ in the extent users are allowed to select the destination. At one extreme, widely adopted in Internet, the user can freely select and change the destination whenever he wants, and without extra charge. The other extreme is that the destination is always the same during a service contract—an example of this approach is a point-to-point, leased-line service. There is an apparent possibility for differentiation from the freedom to select the destination without considering anything particular to a strictly defined scope of destinations, and there are several levels between these approaches.

Although the freedom to select the destination is an important property, it is of minor concern if there is not enough transmission capacity available for the user's purpose. This capacity issue consists of two aspects: what the user is allowed to send into the network, and the quality level that the network can provide. Here is again a good opportunity for differentiation: Users may have significantly different needs concerning both the bit rate and other quality parameters. Consequently, better service can mean freedom to select and change the destination, bit rate, or quality. The counterpart of the freedom is that it makes the network operation and management more difficult.

Two of these aspects, destination and bit rate (or traffic load), are illustrated in Figure 4.4 (from the operator point of view). The main effect of both freedom to change destination and freedom to change bit rate is that the predictability of traffic load inside the networks becomes more difficult.

Predictability of traffic load sent into a network means essentially the accuracy of information about the load in a boundary node related to the future load situation. That is, 100% means that the bit rate is exactly known far into the future—a requisite that very seldom can be met in reality. The lowest predictability means that traffic load is totally unpredictable, something like self-similar traffic with high variability on all timescales. The middle level of predictability may correspond to a traffic process with relatively small traffic variations.

Figure 4.4 Predictability of traffic and destination.



The high predictability could either be inherent due to the original traffic characteristics, or it could be attained by traffic-control actions in boundary nodes. It seems that in the Internet environment, high predictability can be attained only by traffic control.

Predictability of destination means how much information about the destination and route of the flow is available for traffic-control purposes. A figure of 100% means that the route to the destination is exactly known; a low predictability means that there is not much reliable information about the destinations. Although there are inherent differences in the extent of scope between end users, the network can "predict" the destination with high probability only if the number of permissible destinations is small.

4.4.2 *Service Models*

This section briefly outlines some possible service models, marked in Figure 4.4 from A1 to C3. In model A1, a CBR connection is established to a fixed destination for a long duration. This service model requires that the traffic sent to every destination be tightly controlled at the boundary node in a way that no excess traffic is allowed (that is, the service provider applies the guaranteed-service model described in "Implementation of Integrated Services" in Chapter 2). This kind of approach is possible if the bit-rate and destination changes are so rare that the capacity can be updated by the network-management system, or if the changes are more frequent, a signaling system is needed to inform all intermediate nodes about all changes.

The main difference between A2 and A1 is that there are traffic variations that cannot be informed effectively to the intermediate nodes, either because of quickness of the changes or limited capability to transfer load information through the network. An example could be a network with a full-meshed topology and with controlled load service. Because the

service is based on predefined traffic parameters, there are some inherent limits of traffic load, and, consequently, the network can predict the traffic load to a certain extent, but not completely.

In the service model A3, the destination and path are fixed, but the user is allowed to send as much traffic as needed and to change the bit rate whenever he wants. Because the access rate always has an upper limit, the main difference between A3 and A2 is that the average load sent by users is much less than the maximum rate, but occasionally the user is using the whole capacity of the access link. This service could be something like best effort to a fixed destination; as concluded later in this chapter, however, this model does not seem to be feasible from the operator point of view.

The service models B1, B2, and B3 are basically the same as A1, A2, and A3, respectively, with the exception that the traffic destination is not totally fixed, but consists of a group of allowed destinations. Within this virtual network, it is possible to provide all kinds of services: guaranteed, controlled load, and best effort. It should be noted that if the destination of individual connections is fixed, the service belongs to A1, A2, or A3 in this methodology. Therefore, the middle group in the figure is not a synonym for virtual private network (VPN).

In service model B1, suppose that the traffic flow from a source is both constant and strictly policed, but that the destination may change within the virtual network. Although this approach may sound reasonable, the assumption that traffic is constant and strictly policed usually entails that the quality is highly assured. These requirements together with unknown destinations make network dimensioning quite difficult unless it is possible to keep the load level low. Even though B1 is not a reasonable model as the only service in the network, it might be used with some other services that can better exploit the network resources. Service model B2 can be somewhat better in this respect, and the remaining capacity can be left for best-effort service (B3).

In general, it seems that from the traffic-management viewpoint it is of minor significance whether the number of allowed destinations is 10 or 10 million. As a result, the previous consideration of B1, B2, and B3 are largely valid with models C1, C2, and C3 as well.

The fundamental problem to be solved—if the operator tries to maintain high utilization—is how the relevant information is transmitted to all nodes along the path. The accuracy of traffic prediction inside the network depends, therefore, significantly on the capability of the network to adapt to permanent or semi-permanent traffic changes, in addition to the traffic characteristics that makes it difficult to predict the traffic at the edge of the network. In the following pages, a mathematical model is used to evaluate this issue.

4.4.3 A Model for Evaluating Statistical Multiplexing

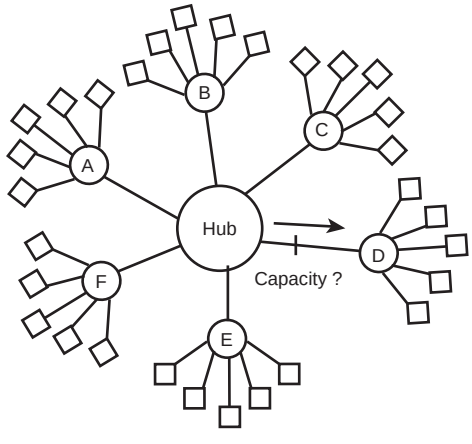
For those who prefer mathematical formulation, the preceding service model discussion is illustrated in the following pages by aid of a relatively simple mathematical model. It should be stressed, however, that the following model, although mathematical in form, is a simplified depiction of a convoluted phenomenon. Therefore, the model should be considered as a mathematical illustration of the service model presented in the preceding section rather than a strict mathematical reasoning. The evaluation is relatively long, but some of the conclusions are so fundamental that only a brief glance will be necessary for most readers. Moreover, numerous figures are used to make the reasoning comprehensible even without considering the mathematical formulae.

The main target of this investigation is to assess the effect of different service models on the efficiency of statistical multiplexing or, in other words, the attainable utilization level. It is always important to keep in mind that there are other, possibly more important targets in addition to efficiency, however, such as fairness, versatility, and robustness.

The Target of Modeling

Consider the network illustrated in Figure 4.5. Your task is to dimension the link from the central hub to one of the other nodes. What you know for certain is the load situation at the beginning of a dimensioning period (t=0). Based on this information, you estimate the capacity required at the end of the period (t=T) that provides sufficient QoS level.

Figure 4.5 A network model.



Basically, the length of the period (T) is the time needed to update the capacity. In case of a completely switched network with an efficient signaling system, the entity under dimensioning is the link capacity needed to support the connection. If it is not possible to alter the capacity reservation during the connection, the length of the period is a typical duration of a connection (for instance, some minutes). In a very efficient system where every change in required capacity of every connection is transmitted through the network, the length of the prediction period could be seconds. In the case of a virtual private network managed through an O&M center, the period could be some days (the maximum time needed to update link capacities). Finally, if the only way to manage network capacity is to increase physical capacity, the length of the period could be months rather than days, minutes, or seconds.

Nevertheless, the most significant issue concerning this simplified dimensioning model is how well it is possible to predict the load situation at the end of the period (t=T) provided that it is known at the beginning of the period (t=0). One of the principal questions is the meaning of load situation. Load level is relatively easy to define for an arbitrary instant of time: The load could be said to be the average bit rate over a period that is needed to empty a full buffer. The reasoning behind this definition is that the buffer can filter variations that occur at shorter timescales, whereas there is no guarantee that the buffer can filter longer variations. Therefore, this example attempts to predict the bit-rate distribution in a timescale of milliseconds or tens of milliseconds.

Estimation of the Load Situation at the Beginning

You are probably wondering how to know the load situation at the beginning of the period. The fundamental problem here is that mathematical models usually require that all parameters are somehow known in advance. You can use as a basic assumption that the load at the beginning is the momentary load at a certain point of time. This is a somewhat pessimistic assumption, because a longer measurement can provide more information about the actual load situation.

To keep the calculations tractable, you can further suppose that all connections are similar on / off sources. In that case, the load situation at the beginning of the period can be presented by two parameters: the total load, M(0); and the capacity used by one flow, B (kbps). Respectively, the load situation at the end of a period can be described by a conditional probability distribution $Pr(M(T)=x|M(0))$. If you can attain a reasonable estimate for this probability distribution, you can decide how much extra capacity is required.

Dimensioning Principle

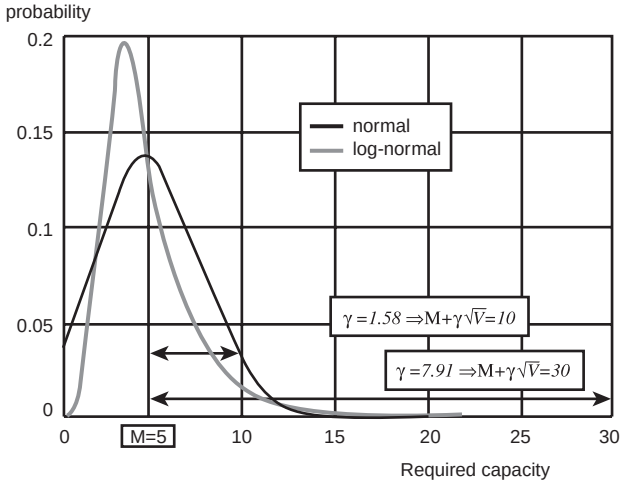
To avoid the tedious calculations of complex probability distributions, you can suppose that the needed capacity (C) depends only on the average (M) and variance (V) of the distribution, as shown in Formula 4.1.

Formula 4.1

$$C = M + g*V^{0.5}$$

The principle of Formula 4.1 is illustrated in Figure 4.6. Although this issue is discussed more thoroughly in section 5.4.3, "Network Dimensioning," in Chapter 5, "Differentiation of Customer Service," some basic explanations are worthwhile here. First, basically any probability distribution with two or more free parameters can be fitted to given mean and variance, and there is no way to determine what is exactly the right distribution in this case or for the intended purposes.

Figure 4.6 Dimensioning rule based on mean (M) and variance (V) of a prediction distribution.



A normal distribution is a natural starting point, supposing that the distribution is formed as a sum of a large number of independent random variables. Formula 4.1 has a very clear interpretation in case of normal distribution: For any given factor g, there is a corresponding probability that the required bit rate exceeds the capacity C. If the capacity is sufficient with probability of 99.9%, for example, you select g = 3.09 regardless of the mean and variance of the distribution.

You may be inclined to suppose that this is exactly the situation in this case: There are usually a large number of more or less independent flows. Unfortunately, this is only a small

part of a complicated issue. A significant part of the inaccuracy (that is, the variance of the distribution) does not stem from the variations in individual flows but from the general uncertainty concerning load levels, traffic processes, and so on. You have no specific reason to suppose that this kind of uncertainty can be modeled by normal distribution in particular when the deviation of the distribution is large compared with the mean value. In that case, log-normal distribution (shown also in Figure 4.6) could be a more realistic choice. In some cases, the results provided by normal and log-normal, or some other distribution, differ remarkably although you keep mean, variance, and exceeding probability the same. In summary, you can apply Formula 4.1; but be aware of its limitations.

Estimation of Variance at the End of the Period

If you suppose that there is no systematic change in load situation during the prediction period, the expected load M(T) at the end of the period is equal to the measured load at the beginning of the period—that is, M(T) = M(0).

The primary question now concerns the variance of the distribution. The variance depends crucially on the probability that a connection is active at the end of the period on the condition that it was active at the beginning of the period. If the period is longer than the average duration of an activity period, this probability is approximately the same as the probability of activity of a customer, denoted by `a`. (Note that this straightforward reasoning is valid only in a homogeneous case.)

Based on this probability, you can make a rough estimation, as shown in Formula 4.2. If you knew the exact value of `a`, the size of population `N` (the number of customers in all other nodes), and the bit rate needed by an active connection, `B`, the variance of the distribution can be calculated from a binomial distribution.

Formula 4.2

$$V = (1-a)* a*N*B^2$$

It should be stressed that all the parameters, `N`, `a`, and `B`, are theoretical measures that can be known only in practical situations and, moreover, Formula 4.2 is based on a fictive case that is much more homogeneous than what could be expected in reality. Because of these reasons, a realistic variance is likely to be larger than what Formula 4.2 gives. In the following evaluation, an extra factor of `e` is used to take this into account (where `e` is usually larger than 1).

If you further take into account that M(0) can be used as an estimate for the product `a*N*B`, you obtain the estimation for the variance of load distribution, as shown in Formula 4.3, at the end of the period on the condition that the load were M(0) at the beginning of the period.

Formula 4.3

```
V(M(T)¦M(0)) = e*M(0)*(1-a)*B
```

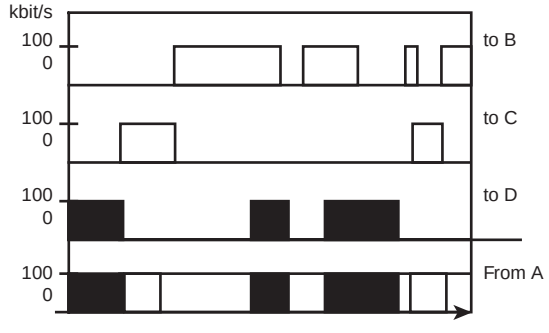
It should be stressed that Formula 4.3 is a rough estimation for the inaccuracy of the traffic prediction, and it should not be considered as an exact method to calculate the variance of any concrete distribution. In practical cases, several factors can either deteriorate or improve the accuracy of the prediction. In particular, parameters `a` and `B` are usually unknown and not the same for all users.

Effect of Destinations Changes

From the viewpoint of one link, it is not important whether the reason for the certain value of `a` is that an active user becomes idle or whether the user remains active but changes the destination in a way that the connection changes to another link. Consequently, the final value of `a`, from the perspective of the link between the hub and node D, depends both on the activity of the customer at the boundary node (a_b) and on the perseverance of destination (a_d). It should be noted that parameter a_b may cover both traffic variations during a connection (whenever the term is applicable), and variations due to the beginnings and endings of connections.

Figure 4.7 illustrates the situation. Parameter $a_b = 1$—that is, the bite rate at the boundary node is constant—and $a_d = 1/3$—that is, traffic is evenly distributed among nodes B, C, and D. Now from the viewpoint of the last link to node D, it is insignificant how a customer attached to node A uses the idle periods that occur on the link to D: she may be either totally idle or she may always send traffic to another node. This separation of a_b and a_d is necessary because there is an essential difference between these two factors at the boundary node, although from the inner link viewpoint they are usually not distinguishable.

Figure 4.7 Traffic from node A and to three nodes B, C, and D.



In the estimation of variance (see Formula 4.3), it was required that the connection be both active at the end of the period and remain on the same link as at the beginning of the period. If these two occurrences are independent, you obtain a simple estimation for a, as shown in Formula 4.4.

Formula 4.4

$$a = ab*ad$$

Finally, if you combine Formulas 4.1, 4.3, and 4.4, you obtain an estimation for the needed capacity at the end of the period, as shown in Formula 4.5.

Formula 4.5

$$C(T)|M(0) = M(0) + g*[e*(1-a_b*a_d)*B*M(0)]^{0.5}$$
$$= M(0)*\{1 + g*[e*(1-a_b*a_d)*B/M(0)]^{0.5}\}$$

Although this example has made a lot of simplifying assumptions, you may be able to draw some preliminary conclusions:

- It is possible to attain high load level only if both destination and traffic variations are under tight control. That is, no extra capacity is necessary if and only if $a_b=a_d=1$ (expect, if you are a clairvoyant, with $e<1$). Although this is an apparent conclusion, and it is somehow an attractive idea to apply it in reality, notice that these requirements can hardly be achieved without significantly reducing the attractiveness of the service.
- The most crucial issue in Formula 4.5 is the B to M(0) ratio—that is, what is the share of one individual connection of the total traffic load? If this ratio is large, efficient statistical multiplexing is not possible. (This is a commonly known phenomenon of statistical multiplexing.)
- If a_b is small, it is practically useless to control the destination of the connections—that is, to increase a_d!
- Respectively, if the destination is not controlled at the boundary node, and destinations vary significantly, it may seem that there is no benefit to controlling the amount of traffic variations (burstiness)—that is, a_b.

The last conclusion is somewhat misleading, however, because a change of a_b evidently affects the momentary capacity needed by a flow (B). Actually if you keep N and M(0) fixed, it is possible to present Formula 4.5 in a somewhat different form, as shown in Formula 4.6.

Formula 4.6

$$C(T)¦M(0) = M(0)*\{1 + g*[\ e*(1-\ a_b*a_d)/(\ a_b*N)]^{0.5}\}$$

Now even if a_d is 0, a decrease of a_b may have a significant positive effect on the allowed load level.

4.4.4 Improving Statistical Multiplexing

The target of this chapter is to evaluate how an operator can improve network utilization by traffic-management actions. The available possibilities are traffic shaping, limitation of destinations, improved knowledge about traffic processes, resource reservation, and quality differentiation. The evaluation is based on Formula 4.6.

Definition of Starting point

The usual starting point of modeling is to fix the necessary parameters—that is, a_b, a_d, N, and B in this case. Now take a somewhat different approach and suppose that you have the following knowledge:

- Current load: $M(0)$ = 5Mbps
- The number of users in a node: N = 500
- Based on practical experience, the average load level should be below 20% of the link capacity to allow sufficient quality when there is no limitation of destination or traffic burstiness, and when some of the flows need high quality.

Formula 4.6 has basically four free parameters: a_b, a_d, e, and g. Because only the production of $e*g^2$ is significant (not the individual values of e and g), however, you can give e a reasonable value (for instance 2). Then you can choice parameter a_d according to the general insight about the probability that the destination remains the same over a relatively long period even though there is no actual restrictions to change it. A guess of 0.25 could be sensible for this parameter. (On the one hand, it is probably not near 1; on the other hand, however, every user probably uses a couple of destinations much more frequently than all other destinations.)

As to the two other parameters, the third assumption of a 20% load level means that either a_b should be smaller or g should be larger than one would primarily expect. As a reasonable compromise, you can suppose that a_b is 0.005 and g is 4.48. This starting situation could be valid with the best-effort service model with few or no restrictions concerning the traffic pattern at the edge of the network or the destination of packets.

If you are dissatisfied with the load level of your network, you basically have five possibilities to improve the situation:

1. *Limit variations in the traffic process at the edge of the network.* For instance, limit the peak bit rate of every user. In mathematical terms, you can increase a_b while keeping $M(0)$, a_d, and N constant.

2. *Limit the transitions from link to link during the prediction period.* Mathematically this means that you increase the value of a_d.
3. *Keep the traffic as such intact, but improve the traffic prediction.* Acquire a more advanced management system that provides better knowledge of traffic processes. Mathematically this means decreasing factor e.
4. *Shorten the prediction period.* This may have an effect on parameters a_b, a_d, and e and by that means improve network efficiency.
5. *Decrease factor g.* You must be aware that the service quality may deteriorate.

As to the context of Differentiated Services, items 1 and 2 seem to be promising because they do not require any changes in the core network. Item 3 is possible as well, but it can only have quite a limited effect if nothing else is done. Item 4 is somehow a natural way to improve the network efficiency, and is widely used (for instance, in ATM networks). It seems that to be really effective, however, you need a resource reservation for every individual flow, which makes this approach expensive to implement. Finally, although item 5 is not necessarily a good idea as such, it can be applied if the total traffic can be divided into several classes with different quality requirements.

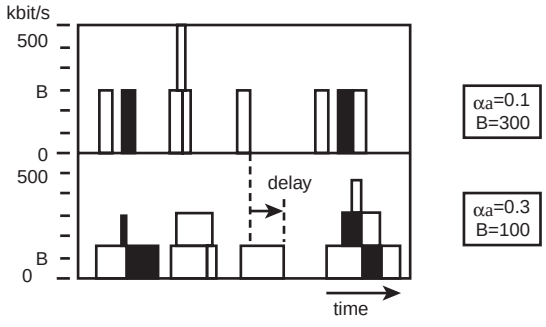
Next you can check which kind of improvements are attainable by any of the preceding items in the case of the numeric example. The figures given previously try to illustrate the reasonable limits of every approach.

Traffic Shaping

If you apply Formula 4.6, a moderate change of a_b from 0.005 to 0.01 improves the load figure to 0.26, whereas a considerable change to 0.05 improves the load figure from 0.20 to 0.44. It should be noted that in the latter case, the length of every on-period is 10 times longer than in the original case, which causes a corresponding increase of delay. Figure 4.8 illustrates this phenomenon.

The acceptability of increased delay depends crucially on the original length of on-period and on the application. With a typical data application, it could be acceptable to decrease the traffic burstiness at the expense of additional delay—for instance, if the effect is that you get a Web page in 3 seconds rather than 2 seconds. On average, however, a higher change than that from 0.005 to 0.05 is likely to result in serious deterioration to quality of service.

Figure 4.8 The effect of traffic shaping.



Destination Limitation

A change of a_d from 0.25 to 1 improves the original load level by a tiny amount of 0.15%. What actually happens, however, if a_d is increased to 1, but everything else is kept constant? As regards the mathematical model, that means that the traffic process at the ingress node is kept intact, but the destination of packets coming from one customer is always the same. You keep the average load on every link and the average number of customers per link unchanged. (This is an apparently somewhat theoretical assumption.)

According to the model, if all other facts remain the same but customers are not allowed to change their destinations, the possible increase of load level seems to be negligible. Only if something else is done as well, would it be worth limiting the destinations. You can, for instance, check the effect of a_d when a_b is already increased to 0.05. The result is that allowed load level can be increased by 1% if a_d is changed from 0.25 to 1—not a remarkable achievement.

The overall benefits are quite small compared to the disadvantages of having strong limits for packet destinations. Under normal statistical assumptions (in particular, destination changes are statistically independent of each other), and with typical Internet traffic, limiting the destinations does not seem to be a practical approach. There can always be other reasons to limit the destination of packets belonging to a certain customer, such as security, but those issues are beyond the scope of this evaluation.

If you make a more realistic assumption in which restricting the possibility of destinations decreases the load, you get even worse results. In the worst case—namely, when the load decreases by the same factor as a_d increases—the allowed load decreases down to 11% because of the impaired statistical multiplexing.

As you can see, this is a complicated issue. If the only service model is a pure constant bit rate without any variability in the load generated by any source, destinations definitely are

of real importance. If there are exactly 50 CBR connections in the ingress node and the destination is selected randomly from two possibilities ($a_d = 0.5$), for example, Formula 4.6 gives an allowed load of 0.61. Moreover, even if the destination is selected from a huge amount of different possibilities, the allowed load is as high as 0.52. Therefore, it seems that the destination should be limited totally, if you are pursuing high utilization only, or not at all, if you prefer a more attractive service model with freedom to select destinations.

Therefore, in the case of CBR traffic, it is useful to strictly limit the destinations to attain higher utilization. This inference could be valid in general only if two conditions are true at the same time: (1) CBR traffic forms the majority of the traffic, and (2) destination limits do not essentially deteriorate the attractiveness of the service. It seems quite improbable, however, that both conditions can be fulfilled on the Internet.

Knowledge Improvement

A change of e from 2 to 1 improves the load figure from 0.20 to 0.33. In the case of the Internet, however, it is likely that factor e can be decreased by 50% only with very laborious effort (if ever) because of the intrinsic unpredictability of Internet traffic. The main difficulty of traffic prediction is related to two inherent control loops:

- The bit-rate regulation of TCP/IP is based on the number of lost packets—that is, on the load level inside the network.
- The decisions made by end users depend on the perceived quality. For instance, the time needed to download a figure through the Internet may have a significant effect on the number of figures to be downloaded later.

These issues together make it extremely difficult to evaluate Internet traffic with any mathematical model. Therefore, the possibility to essentially improve the accuracy of traffic prediction may require a huge effort, including modeling of human behavior. See, for instance, "Where Mathematics Meets the Internet," which discusses the difficulties of understanding Internet traffic (Paxson and Willinger 1998).

Resource Reservation

It is important to consider this approach more thoroughly, because it seems that we very easily return to this model even though it is somewhat opposite to the basic philosophy of Differentiated Services.

If a resource reservation system is applied to every individual flow, you can fix the destination—that is, increase a_d up to 1 and decrease e down to 1, and perhaps increase a_b from 0.005 to 0.5. This means that a customer is supposed to be totally idle 99% of the time

and during the 1% of the time he has a connection to a certain destination, and the connection is active 50% of the time. But then there are only five active connections at the beginning of the period, and five is so small a number that there is no room for any realistic statistical multiplexing, and therefore, the average load cannot be higher than a_d (0.50).

That is a notable improvement compared to the original value of 0.20. There is, however, a big concern related to this seemingly good result. Although this "packet-layer" utilization can be high, this approach introduces an additional "connection layer." If there is no change in the actual traffic process, an additional control layer may deteriorate the whole situation rather than improve it.

The basic reason is that you cannot expect that exactly five users are needing a connection at all times. The traditional way to evaluate this situation is to apply the Erlang blocking formula discussed in section 5.4.3, "Network Dimensioning," in Chapter 5. Basically, the Erlang formula gives the probability that a call attempt will fail due to lack of resources. If the Erlang blocking formula and the 0.1% call-blocking standard is applied, you must reserve room for 14 contemporary connections. The final result is an average load of 0.18. Moreover, this is the result even though you supposed that the average load (five connections) was exactly known.

In short, if and when there is intrinsic uncertainty in the future traffic demand, an additional control layer cannot solve the network-dimensioning problem. In addition, because a service model with a predefined bit rate and destination could be less attractive, the outcome is not very promising from an efficiency point of view. Yet, there certainly are some clear benefits, such as more advanced pricing and the fact that resource reservation favors existing connections over new connection requests. (This issue is discussed in section 4.3.2, "Pricing Models and Predictability of Quality," earlier in this chapter.)

As to the pricing, although it is promising way to smooth traffic (that is, to increase a_b), it can be argued that you can use a similar tariff structure without any reservations inside the network. If the favoring of existing flows is an important service characteristic, and the probability of a conflict is not very small, resource reservation could be a reasonable solution. On the contrary, if the probability of a conflict is negligible, the actual result apparently is similar both with and without reservations. The following example, "CBR Service Versus DiffServ," illustrates the situation.

CBR Service Versus DiffServ

Assume that the fictional service provider Fairprofit wants to offer guaranteed services based on constant bit-rate connections, mainly for real-time applications. Fairprofit defines the customer services in a way that users are not allowed to exceed a bit-rate limit even occasionally. The service provider assumes that this service is used only by applications that really need high quality.

Under these assumptions, customers always have to select the bit rate in advance, although it is possible that they do not exactly know the necessary bit rate. Therefore, customers usually need to reserve more capacity than what they need on average. From the customer viewpoint, there seems to be two practical solutions:

- The user reserves extra capacity in addition to the best prediction of the required mean bit rate.
- The user starts with a bit-rate level somewhat less than what he expects to be necessary, and then increases the reserved bit rate until the quality is sufficient. The need and meaning of a guarantee for the user in this case is that when a sufficient level is reached, it remains sufficient for a relatively long period.

In both cases, the reserved capacity is more than the average bit rate. You can apply Formula 4.1 to approximate the relationship between required capacity (C), mean bit rate (M), and variance of the bit-rate distribution (V). Assume the following numeric values for an individual user:

$M = 100\text{kbps}$

$V = (10\text{kbps})^2$

$g = 4$

As a result, the bit-rate reservation for one flow is 140kbps, and according to normal distribution the probability that 140kbps is not sufficient is approximately $3*10^{-5}$. If the total capacity available for this service class is 10Mbps, there can be altogether 71 connections at the same time, if you suppose that Fairprofit offers a true CBR service without statistical multiplexing on the packet level.

From customer viewpoint, the availability of quality is defined by the probability that the connection request is accepted. That probability can be estimated by the Erlang blocking formula as a function of average load level. Figure 4.9 shows the result. Note that an average load level of 7.1Mbps means that the offered load on the connection level is almost 100%.

The service provider may determine the allowed load level based on this curve, and on the target value for availability of quality. For instance, availability of 99.9% is achieved by an average load level of 5Mbps. (The corresponding average reserved capacity is 7Mbps.)

This could be acceptable for Fairprofit, provided the total cost of the system is not too high. However, the CBR service with reservations for every connection is somewhat hard to implement and manage. Therefore, Fairprofit may be interested in other options as well. The other alternative provided by Differentiated Services is that Fairprofit keeps the service model exactly the same (in particular, user are paying for a virtual reservation the same tariff as for a real reservation), but accepts all connections and even all packets into the network.

What then is the availability of quality as a function of load level? You can start with the fact that bit-rate variations occur because of two basic reasons: connection-level variations, and packet-level variations of individual connections. You can assume that enough quality is available for a connection if (and only if) the momentary load level both at the beginning and at the end of the connection is low enough to make certain that quality is sufficient.

You can estimate the sufficiency by a simple algorithm: If $\{M + 4*V^{0.5} < 10\text{Mbps}\}$, then there is enough capacity; otherwise the quality is supposed to be unavailable for all connections. The result of this model is

also shown in Figure 4.9. Although this model is a rough evaluation of a complicated issue, it may allow some practical conclusions.

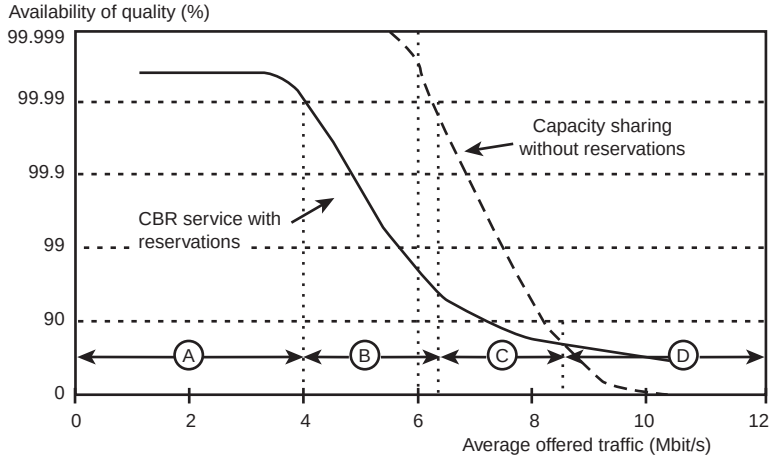
From the service provider viewpoint, there are four load regions:

- *Region A*: In region A (from 0 to 4.4Mbps in the example), both systems provide excellent performance—that is, higher availability than 99.99%.
- *Region B*: In region B, the performance of the reservation system is deteriorated because some connection requests are discarded, whereas the sharing system still offers excellent performance.
- *Region C*: In region C, both systems yield deteriorated performance, but the sharing system still provides better performance than the reservation system.
- *Region D*: In region D, the reservation system works better because it can provide good service for some of the connections; equal sharing, on the other hand, leads (at least in theory) to a situation where no one can obtain sufficient quality.

Fairprofit may conclude that the reservation system is not the optimal one, if the network-management system can somehow limit the possibility of an overload situation.

Moreover, Fairprofit can improve the overall performance of both systems. The reservation system can be enhanced by applying statistical multiplexing by applying controlled load service rather than guaranteed service (or VBR rather than CBR in ATM networks). Capacity sharing can be improved by priorities that can guarantee high quality for some flows even during overload situation—that is, by Differentiated Services.

Figure 4.9 Availability of quality with and without reservations.



Quality Differentiation

Suppose, for example, that 20% of the traffic really needs high quality, and 80% can cope with a moderate quality. If you have a tool to protect high-quality traffic from the harmful effects of low-quality traffic, you can significantly increase the load level.

Suppose further that high-quality traffic requires a safety factor of 8. (As high a value as 8 could be required mainly because you cannot be sure that normal distribution is valid in real cases.) Therefore, higher-quality traffic can attain a load level of 0.077, if the other traffic parameters are kept the same, except that a_b is changed to 0.001. Consequently, the total load could be 5*0.077 = 0.385. This is quite a conservative evaluation, because high-quality connections are usually controlled more tightly than typical Internet traffic, which makes it possible to reduce the prediction variance.

If you suppose that the other part of the service is pure best effort, you can suppose that TCP or some other protocol adjusts the bit rates if necessary. Thus a quite low safety factor could be acceptable (for instance, 1.5). According to Formula 4.6, the total load could be approximately 0.42. Because the first limit is tighter, the total load could be enhanced from 0.20 to 0.385 purely because of the service differentiation. Note, however, that this improvement is possible only at the expense of decreased quality of service for some customers.

4.4.5 *Conclusions About Statistical Multiplexing*

From an efficiency point of view, the following conclusions are reasonable:

- Traffic shaping is recommendable when it is feasible, taking into account the possible reduction of service attractiveness.
- The real benefit of destination limitation is unclear, but it seems to be usually small.
- Improvement knowledge about traffic processes should always be used when there is a reasonable possibility without too high a cost.
- Resource reservation does not usually improve the network utilization; the benefits are elsewhere, such as in the better predictability of quality.
- Traffic classification surely provides promising results.

The concrete objectives for Differentiated Services based on this mathematical evaluation are as follows:

- Network service should encourage traffic shaping made either by the applications in customer equipment or by the network operator at the network edges. Encouragement means that the customer benefits from smoother traffic either by obtaining lower price, lower packet-loss ratio, higher throughput, or shorter transmission delay.
- Service differentiation is the key tool to improve cost efficiency.

4.5 Traffic Handling

Although network utilization is an important target when designing a network, it is definitely not the only one. This section discusses some other goals that are not directly related to network utilization, but are still important for the overall usability of the network service.

The two main aspects of traffic handling are *urgency* and *importance*. These terms may need some clarification, because a lot of similar terms are used in the literature. All combinations of urgency and importance are possible: A packet can be urgent and important, urgent but not important, important but not urgent, or not urgent and not important. In practice, high urgency means that a packet should be delivered as quickly as possible—in other words, with as small delay as possible. Thus urgency is a characteristic needed by a packet, and delay is a characteristic of packet handling.

The term *importance* is used here rather than the more technical terms *drop precedence* or *drop preference* because of two reasons:

- Drop precedence is somewhat troublesome to use because higher drop precedence means lower service class (or lower priority if you prefer that term).
- Importance is a more general term. Although importance of a packet and drop precedence can often be used as synonyms, drop precedence is not necessarily the only way to implement importance differences between packets.

In addition to these aspects, packet handling may take into account some other issues, such as bandwidth division, routing, and support for adaptive applications. The main issue from the Differentiated Services viewpoint is whether any of these issues requires special packet handling in interior nodes, or has any significant effect on the implementation of a Differentiated Services network.

Note

In this book, importance level refers to information about the relative importance of a packet to be used for traffic management.

4.5.1 Urgency

From an application point of view, two main requirements for network service are that available bit rate is high enough and that packet-loss ratio is low enough. For some applications, however, these are not sufficient requirements because the packets should be delivered within a certain period of time; otherwise, the packet is not more useful than a totally

lost packet. This requirement can be called *urgency*. It seems that with the current level of network performance, a special urgency service is needed only for interactive applications, such as IP telephony or videoconferencing.

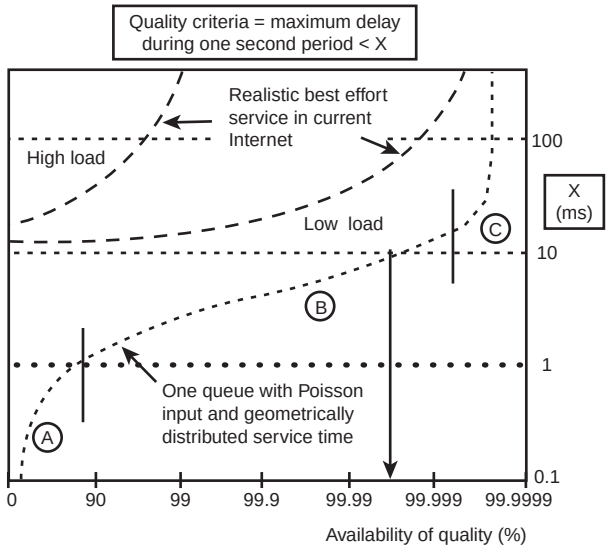
The rate at which an IP telephone generates packets is usually constant. These packets are played back at the same constant rate to make the audio signal comprehensible. Basically, there are two ways to realize this: (1) to design the network service in a way that the delay through the network is strictly constant; and (2) to use a buffer in the receiving equipment in a way that the packets can be used at the same rate as they were generated. A telephone network is based on the first option, but packet networks are usually based on the second option.

Even if you take into account the effect of playback buffers, however, there is a clear need to control delays inside the network. Although the unavoidable delay components (packetization and transmission delays) are out of the network's control, queuing delay is the controllable part. Because the basic unit for delay analysis is one buffer, it is possible to make a rough estimation for the availability of quality in one buffer. The result could be something like presented in Figure 4.9 by the S-shaped curve. For instance, a delay criterion of 10 milliseconds may theoretically yield availability of 99.993%.

With a relatively high probability (for instance, 70%), the queue is empty or almost empty and the packets do not encounter any significant queuing delay or delay variation (region A). In the middle region (B), the queuing delay increases, and a high availability can be reached if the quality requirement is increased and traffic load is under control. Finally, with certain probability, the instantaneous load level becomes too high in a way that even a very large buffer cannot prevent packet losses (region C).

This type of curve can be obtained in one buffer and controlled circumstances in which the load levels are relatively stable. The reality is again much more complex. In particular, if all flows share the same service (best-effort service in the current Internet), a majority of the traffic is likely using TCP. One straight consequence is that delay variation is usually high. Figure 4.10 presents two likely scenarios: one with a high load level and another with low load level. Although the picture does not represent any real measured situation, it seems unlikely that any tight delay requirement can be obtained with one service class unless the average load level is very low.

Figure 4.10 Availability of quality with delay criterion.



4.5.2 *Importance*

It is crucial to attain some level of consensus about the importance; otherwise, the result could be as in *Alice's Adventures in Wonderland*:

> "Unimportant, of course, I meant," the King hastily said, and went on to himself in an undertone, "important—unimportant—unimportant—important—" as if he were trying which word sounded best. Some of the jury wrote it down "important," and some "unimportant" (Carroll [1865] 1970).

If some nodes assess the packets of one aggregate as important and other nodes as unimportant, the final result is a total mess. But why do the nodes, after all, need to know the importance of every packet? To make, when necessary, a reasonable decision as to which packets can and should be discarded if all packets cannot be forwarded.

Several aspects of importance have been discussed in this chapter, such as the differences in availability of quality can be considered as the main target of marking packets less and more important. Now the key issue is whether there is truly several kinds of importance, in a way that the interior nodes have to be aware not only of the level of importance but also of the type of importance.

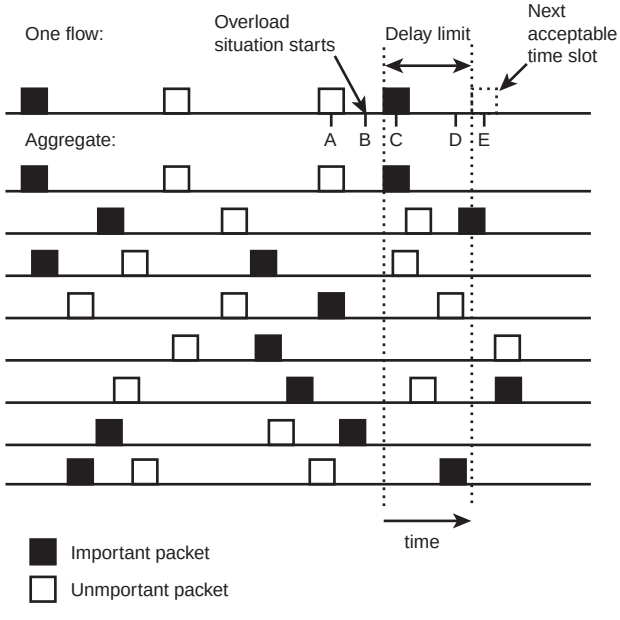
An issue is the need to mark some packets belonging to a flow as more important than some other packets of the same flow. This aspect of importance could perhaps be essentially

different from other importance aspects. Should the interior node make a different decision if the packet is marked important by the customer or by the boundary node? It seems apparent that sometimes this is the case. This is a somewhat misleading viewpoint, however. The right question is, is there a reason why the boundary node cannot map the user desire in the same importance scale as all other relevant issues? From that perspective, it is much harder to identify any need to have two independent importance scales.

Important Versus Less Important

The system promoted in this book is based on the assumption that it is better to map all kinds of importance issues into the same general structure, if possible. In particular, interior nodes should see only one importance dimension, and the boundary nodes should map different importance aspects to that one-dimensional scale. This is done mainly to guarantee a consistent and simple decision-making system inside the network. It may be argued, however, that this is a favorable system also from an individual flow viewpoint, as illustrated in Figure 4.11.

Figure 4.11 Selection of packets to be discarded.



Assume, for the sake of argument, that an individual flow consists of two types of packets: important and less important. The network node encounters a situation where it needs to limit the rate of the flow, because of an overload situation (instant B in Figure 4.11). The preceding packet of the flow has been accepted at instant A because there was no overload situation at that time.

The problem when the node makes the decision individually for every flow is that there are usually not available several packets from which the node can select the least important packet to be discarded. In Figure 4.11, there is only one packet within the time limit when the discarding decision is made. An important packet could perhaps be delayed for some time, but if the next acceptable time slot for the packet is not within the delay limit, the packet has to be discarded. It is easy to say that less important packets are discarded, but in reality there could be only one packet available during the time on which the decision is made. In that case, there is no other possibility than to discard that packet regardless of the importance.

If the discarding decision is made for an aggregate stream, the possibility to make a reasonable decision is much better because there are probably both important and less important packets. As a result, important packets do not need to be discarded. In general, the more packets available, the better the possibility to make a reasonable discarding decision. This is one more reason to avoid unnecessary bandwidth fragmentation.

Urgency Versus Importance

The main question here is whether importance and urgency (or delay requirement) are essentially different things—that is, whether an important packet is always an urgent one, and whether a less important packet is always a less urgent one. As discussed previously, interactive applications create the need for real-time service in IP networks; most of the other applications can cope with the delay characteristics of the current best-effort service.

Therefore, the explicit need is for a network service that provides low delay but relatively high packet-loss ratio. If one of the main targets of importance differences is availability differences, as this book has supposed, then that kind of service could be as reasonable as a service with low delay and low loss ratio. A service with low delay and low importance could be exactly what is needed to provide inexpensive but applicable global IP telephony service. In summary, it is necessary to clearly distinguish *importance* and *urgency* in such a way that operators can freely combine different levels of importance and urgency.

Note

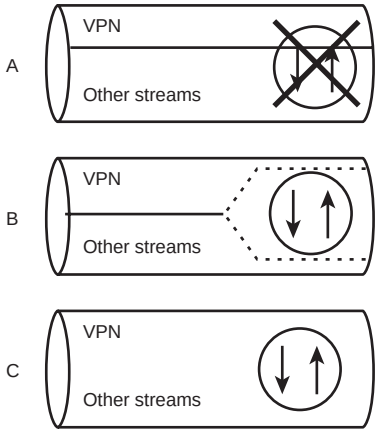
In this book, importance level refers to information about the relative importance of a packet to be used for traffic management.

4.5.3 *Bandwidth Division (Virtual Private Networks)*

A virtual private network (VPN) is something that a large customer or enterprise wants to use for his own purposes. The enterprise typically has multiple locations and wants to interconnect them using an IP backbone network. The VPN is defined by the collection of the locations and traffic-management rules within that VPN.

VPNs can be roughly classified into three categories, as shown in Figure 4.12. In category A, each VPN has its own dedicated resources (bandwidth and buffer space) on every link in the network. On the one hand, the bandwidth is reserved for it all the time regardless of the actual use of the capacity; on the other hand, the flows belonging to the VPN are not allowed to use any spare capacity reserved for other traffic streams in the network. In a way, this scheme is fair and comprehensible, because the customers get exactly what they have paid for. If there are large numbers of parallel VPNs, however, this approach could be very inefficient because the capacity fragmentation may seriously deteriorate the positive effects of statistical multiplexing.

Figure 4.12 VPN categories.



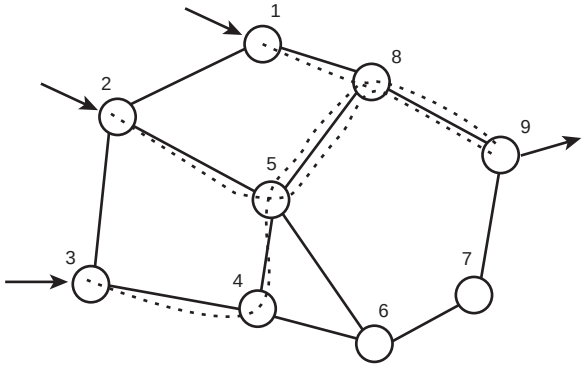
Approach B allows limited statistical multiplexing between different VPNs. The rules of capacity sharing are an essential part of the service contract. It is possible, for instance, to define that a VPN always gets a certain minimum bandwidth on each link regardless of all other traffic streams; and that it can use any free capacity on the link (perhaps up to a certain limit); and that, correspondingly, other streams can use resources unused by the VPN. These rules may also take into account the importance levels of the packets, provided that the service provider has some level of control of the importance marking of the packets.

This kind of system could be more complicated to manage than approach A, and it may induce some fairness problems, if the load levels of different VPNs vary considerably. Nevertheless, the statistical gains are so significant that they presumably exceed the drawbacks.

Approach C is an extreme case of the preceding approach with very flexible statistical multiplexing between VPNs inside the network. Although this approach may look, at first glance, inappropriate, it could be feasible if the packet marking were done carefully in the boundary nodes. It should be noted, in particular, that the packets with highest importance should be discarded with minimal probability. From a packet point of view, it is totally irrelevant whether the reason for successful transmission through the network is capacity reservation or marking to a high importance PHB. If the network is properly dimensioned and the number of high importance packets is kept relatively low, the overall result could be both appropriate from the customer point of view and efficient from the management point of view.

One fundamental matter should still be stressed: Even though the traffic of one VPN can be under strict control in the first link, the control cannot be strict inside the network without a very dynamic provision system (that is, signaling and per-flow reservations). In Figure 4.13, for instance, even though the traffic entering the network in nodes 1, 2, and 3 is within the defined limits, there is a realistic possibility to exceed the VPN capacity in the link between nodes 8 and 9 if all the traffic from 1, 2, and 3 is directed to node 9.

Figure 4.13 A VPN network with a possible conflict.



Four ways to avoid excessive packet loss can be identified:

- *Limit the traffic between every pair of nodes.* The extreme case is a full mesh of constant bit-rate channels between all nodes.

- *Limit the total traffic leaving each node low enough to guarantee that there is no packet losses inside the network.* In an extreme approach, the maximum traffic is the lowest capacity divided by (n–1), where is n is number of nodes.
- *Use adaptive routing and direct some of the traffic through nodes 6 and 7.* This approach is discussed briefly in the next chapter.
- *Use either VPN types B or C, as presented in Figure 4.12.*

Even though the last approach seems to lack the guarantee provided by the three other approaches, the significantly better efficiency makes it a much more attractive solution. But then, once again, you need a systematic differentiation and control of packets and flows to attain an appropriate result that is fair, robust, and versatile.

4.5.4 Routing

As mentioned in the preceding chapter, routing is one tool to alleviate congestion situations. Nevertheless, routing is not discussed extensively in this book because of three reasons. First, the standardization of Differentiated Services mainly concerns packet forwarding, not routing. Second, adaptive routing can only have a limited effect on the overall utilization: It is not necessary in lightly loaded networks, and the effect could be even harmful if the network is highly congested and the routing mechanism is not designed carefully. On the contrary, in the middle region of a moderately loaded network, adaptive routing could be useful.

And finally, in case of a complex PHB structure, the rerouting system must be aware of the relationships among different PHBs. Two PHBs belonging to the same PHB class, for example, should not use different routes. (PHB class is used in the meaning explained in the section titled “Terminology” in Chapter 3, “Differentiated Services Working Group.”) In summary, adaptive routing that takes into account PHB information cannot be the main tool for controlling network traffic or for quality provisioning.

Another issue to consider is that when the network-management system composes the routing table, it can take into account the average load level of different PHBs with different weights. The network operator may decide to maximize the availability of service for the most important PHBs, and then perhaps take into account the expected load levels of some moderately important PHBs with smaller weight, while totally ignoring the lowest level PHBs.

Finally, there are some special cases in which the route may depend on the PHB of the packet, such as satellite links. PHBs with good real-time characteristics should use satellite links with long transmission delay only if there is not other alternative with better delay

characteristics. Also, if a link occasionally has a high bit-error ratio, it can be used for PHBs without high-quality expectations. Although these are possible scenarios, they have no significant effect on the requirements of the PHB architecture.

4.5.5 Support for Adaptive Applications

As already emphasized several times, all systems relying on signaling are mostly beyond the scope of the Differentiated Services standardization and this book. This choice means that it is usually impossible to make any admission control for individual flows, let alone to inform interior nodes about changes in traffic parameters of individual flows. Another approach, in a sense reverse, is that the network informs senders about the available capacity inside the network.

There are two different aspects related to this congestion-notification approach. The first aspect is the need to have a mechanism to inform network nodes that the flow is willing to adjust its bit rates during a congestion situation, and to inform about the actual congestion situation. The current view seems to be that the 6-bit field currently under standardization process is not used for these purposes.

The second aspect is more interesting in this connection: The question is, what is a proper consequence in an interior node if a packet is marked to belong to an adjustable flow? There are basically two approaches. The system is used merely to inform the sender of a congestion situation without any effect on the treatment of the packet. Because this system does not have any effect on the implementation of Differentiated Services inside the network, it is chiefly beyond the scope of this evaluation.

Another possible approach is that the packets obtain better treatment because the flow (or application, or finally customer) promises to reduce the bit rate if the network informs about congestion situation. In reality, this system means that the packets of those flows should have higher importance than corresponding packets belonging to nonadjustable flows. It seems that this system can and should be realized in the boundary nodes, without any need to change the general structure of PHBs (except that an additional importance level could be necessary). One possible way to tackle this problem is that packets with a low importance level could either be sent into the network or dropped, depending on congestion monitoring as suggested in "A One-Bit Feedback Enhanced Differentiated Services Architecture" (Arora *et al.* 1998).

The general conclusion is that the traffic-condition function can take into account information related to congestion notification, but there is no need to modify the PHB system. It should be stressed that a mere promise to take the congestion notification into account cannot be sufficient reason to improve the importance of the packets, because there is an apparent possibility that the customer will mark all flows as adjustable.

4.6 Framework for Per-Hop Behaviors

Now it is time to build the technical framework for designing per-hop behaviors. From the lengthy discussion about importance, delays, availability, and predictability, it is fair to conclude that three fundamental dimensions are necessary for the framework: importance, urgency, and bandwidth. This statement entails that the specification of each per-hop behavior should somehow address each one of the three aspects: What is the importance of the packets? What is the urgency of the packets? And what is the bandwidth that should be allocated to the aggregate (if any)?

4.6.1 Essence of Quality

The first question is, What form is used to define each quality aspect? Basically, there are three possibilities: quantitative, qualitative, and relative. This is the same categorization as presented in section 4.3 of the framework document (Bernet *et al.* 1998). (See section 3.3.4, "A Framework for Differentiated Services," in Chapter 3 of this book.) For instance, with a PHB class that has a moderate packet-loss ratio, provided that this is a proper measure for importance, these service categories can be depicted as follows:

- *Quantitative service*: Packet-loss ratio should be less than 10^{-4}.
- *Qualitative service*: Packet-loss ratio should be moderate, which could mean that it is less 10^{-4} most of the time, but during busy hours it can be higher; and during idle hours it is usually zero.
- *Relative service*: Packet-loss ratio of this PHB should be smaller than that of any PHB with lower importance.

There seems to be a strong temptation to quantify all issues and to suppose that quantification itself makes the system somehow better. To a certain extent, this may be true—for instance, the verification of the system or service provision is easier if there are definite numeric specifications. The operator may explicitly choose to apply that approach. It seems that the framework of Differentiated Services does not provide enough tools for implementing truly quantitative services, however, with the possible exception of one service class with high quality.

Nevertheless, the quantitative service model is still the only acceptable target for many system developers. Therefore, it is highly probable that there will be a lot of effort to direct Differentiated Services toward the quantitative service model.

On the contrary, this book recommends that if a service provider wants to *extensively* apply the quantitative service model, another technology with complete specifications (such as ATM or Integrated Services with RSVP) should be used rather than Differentiated

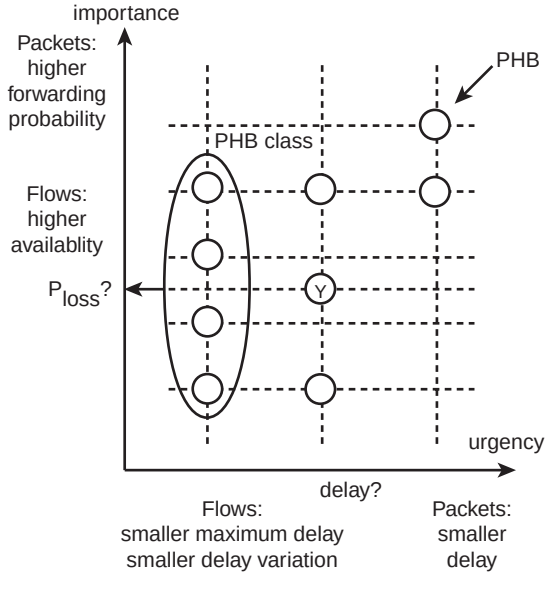
Services. This issue, of course, depends on the exact meaning of quantitative and qualitative. Although quantitative services likely use more advanced control methods with admission control and per-flow reservations, the perceived quality may sometimes be higher with qualitative service than with quantitative service.

This statement leaves two alternatives, the quantitative model and the relative model. Laying aside the numeric formation of the services, the overall structures of these two approaches seem to be similar, because low and high mean essentially the same as lower and higher. In particular, it is hard to distinguish any difference between these two models on the packet-handling level in interior nodes. What are the difference in packet-level implementation of two systems, one consisting of low and high importance classes, and another one consisting of lower and higher importance classes?

4.6.2 *Relative Scales for Importance and Urgency*

Therefore, a system with relative scales can be a good basis for a Differentiated Services framework. Therefore, if a service provider wants to apply qualitative scales, he can bind the scales to certain fixed points by appropriate operation and management functions (but that does not necessarily yield any changes to the actual packet-handling actions). Figure 4.14 shows the result. You can think of this structure as a cobweb, with the capability to stretch and shrink according to the burden without essentially changing its logical structure.

Figure 4.14 Relative scales of importance and urgency.

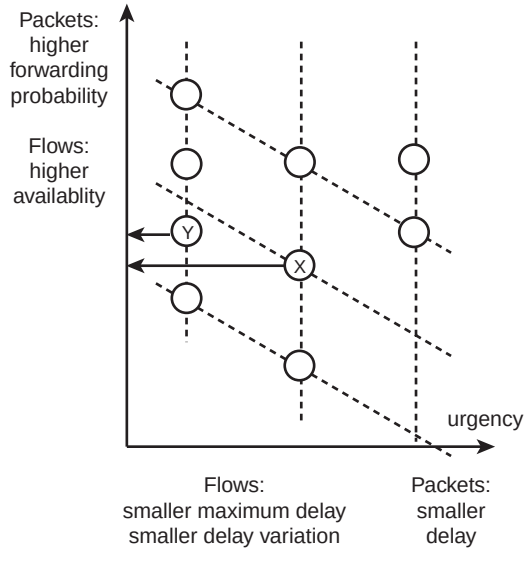


Every PHB has a definite position in relation to all other PHBs: The delay and packet-loss ratio of one PHB are either smaller, higher, or approximately the same as that of another PHB. Therefore, the definite values for delay and packet-loss ratio of a PHB depend on the load level in the network. In a relative system, service providers do not necessarily say much about the expected values of quality parameters.

It is fair to expect that the number of delay classes is limited, probably to not more than four, because each delay class requires its own queue in every node, and a large number of delay classes increases the network-management burden. The delay of individual PHBs belonging to the same PHB class cannot vary significantly, because the service provider is allowed to re-mark packets within a PHB class from a PHB to another one, and different delays may yield packet reordering. Therefore, even though the structure is somewhat elastic, the PHBs within one PHB class should be kept on the same vertical line.

A less clear issue is whether the other scale (importance) should behave in the same way. Figure 4.15 shows a situation in which the importance order of PHBs X and Y is changed from the original situation depicted in Figure 4.14. Whether this kind of change is allowed depends on the overall service model, and on the available mechanisms in network nodes. With certain mechanisms, the packet-loss ratios can be controlled only within a PHB class. On the contrary, with some other scheduling and buffering mechanisms, it is possible to keep the packet-loss ratio of two PHBs belonging to different delay classes approximately the same, even though the loads of the two classes vary in different ways. The question of which one of these approaches is better can be left for the service provider.

Figure 4.15 Changing the order of two PHBs in relation to packet-loss ratio.



The main difference between a relative and a qualitative service system is that in a qualitative system the service provider makes some additional effort to stabilize the quality of PHBs, whereas in a pure relative system the service provider does not make such an effort. The most reasonable way to stabilize the quality levels is to somehow control the offered load level of each PHB. There are apparently a wide variety of possibilities between minimal control (relative service model) and maximal control (qualitative service model).

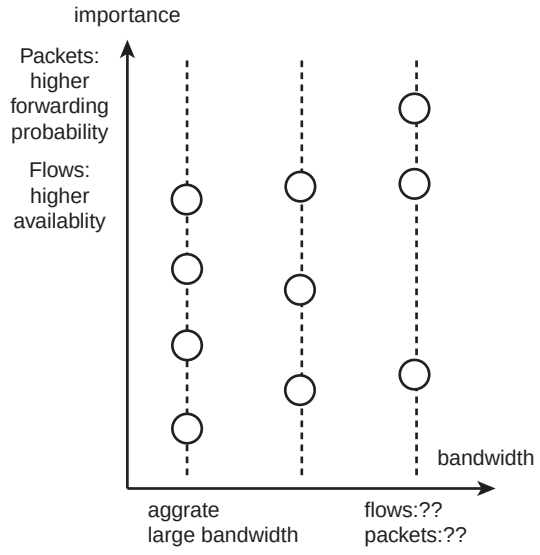
4.6.3 *Relative Scales for Bandwidth*

Although the relative structure shown in Figure 4.15 seems to be valid with importance and delay scales, that it is valid with the bandwidth scale as well is not evident. What do you really want to accomplish with the separate bandwidths? There are definitely different ideas and answers. The main approach adopted in this book is that bandwidth can be a useful tool to separate user groups that inevitably will be separated. That kind of group can be called a VPN group (although the purpose of the separation is not necessarily a VPN).

In this case, a quite natural approach is to apply within each VPN group a similar system, as shown in Figure 4.14. In that case, there is no need to determine the relationships among PHBs in different VPN groups. The main problem of this approach is that because the amount of PHBs is very limited, there cannot be a large number of VPNs within one PHB domain. Of course, some other mechanisms such as MPLS can be used to inform network nodes about the VPN of each packet, but the network nodes must provide appropriate mechanisms for every VPN group (for instance, separate queues for each PHB class in every VPN).

The second possibility, also mentioned in the framework document (Bernet *et al.* 1998), is that each PHB class gets a relative amount of bandwidth in every link (for instance, in proportions of 1, 2, and 4). Each PHB class may then have a number of PHBs with different importance. There is no procedure for determining the relationship between two PHBs belonging to a different PHB class. A more serious concern is that it is impossible to infer much about the information that a PHB group has larger bandwidth than another one, in particular, it does not tell much about the situation from the perspective of an individual flow or individual packet. These issues depend crucially on the traffic-conditioning function in the boundary nodes and the traffic-management functions that regulate the resources given to different aggregates.

Figure 4.16 A PHB system with importance and bandwidth scales.



These three dimensions seem to be sufficient for depicting most of the PHB proposals and most of the integral characteristics of any service based on PHB. Some aspects may need further assessment, however.

4.6.4 *Predictability of Quality*

Based on the evaluation in the first part of chapter, the main concern is the predictability of quality in a short timescale—that is, the need to keep the quality of individual flows more predictable or more constant than is possible in the basic Differentiated Services system. The main application of this kind of system is to give a priority to existing flows during emergence of congestion, in particular to IP telephony calls. One straightforward approach is to make a resource reservation for every individual flow. Because that kind of system is beyond the scope of Differentiated Services, we need another approach.

The essential nature of a system to give priority to existing flows seems to be that the packet of existing flows should have somewhat higher importance than the packet of starting flows. Although it is not clear how to implement this in reality, the required procedures should evidently be done in the boundary nodes. Therefore, this issue seems to require only traffic-conditioning functions without any need to introduce new PHBs.

Summary

This chapter addressed three fundamental issues:

- *Availability of quality*: A tool to compare service models, from guaranteed services to relative services with one universally applicable measure
- *Efficiency of statistical multiplexing*: An evaluation that takes into account predictability of load and destination
- *Predictability of quality*: The need to provide constant quality for existing flows

The common objective of all these considerations has been to provide a framework and tools for fair and extensive evaluation of all service types from pure best-effort service to highly guaranteed service.

The most concrete result of all the considerations is a simple and consistent framework for per-hop behaviors. The framework consists of three dimensions: importance, urgency, and bandwidth. This framework is used extensively in Chapter 7, "Per-Hop Behavior Groups," to evaluate and compare different PHB proposals.

> Thus endeth the afternoon's talk of Raphael Hythoday concerning the laws and institutions of the island of utopia.
>
> —Sir Thomas More