CHAPTER 5

Differentiation of Customer Service

In general, this book strongly emphasizes the wholeness of the system rather than the separate parts of the system. And although the potential forms of Differentiated Services are endless, the most relevant issue is its purpose: All parts of Differentiated Services have to support the building of an attractive and lucrative network service. This objective, closely related to customer service, is the topic of this chapter. Three specific issues are addressed:

- Overall service models
- Ways to request specific service
- Pricing structure

Customer expectations related to different services may vary remarkably. The same users who expect their telephone company to provide consistent high quality may content themselves with variable quality when it comes to the Internet. Various reasons may account for this. One reason may be the different pricing schemes. A customer can usually obtain an Internet service to any destination for a set monthly fee. Telephone calls (except local calls), on the other hand, are usually incrementally priced, increasing with the distance to the destination and the duration of the call.

In the future, as Internet customers begin to exploit new technology to make telephone calls and to use the many other services that the Internet infrastructure makes possible, the expectations of different customers will also change (making those expectations much more difficult to predict, from a service provider's perspective). The general service model, closely related to the marketing of the service, is a key tool to handle this intricate issue. The service provider may promise a guaranteed service with definite quality characteristics, or it may merely sell shares of network resources without any explicit guarantees.

The dynamic nature of service requirements is another critical consideration that providers must keep in mind. A service level can be either permanent (that is, guaranteed) or the provider may allow the customer to inform the network when service requirements change (perhaps even every second). These two levels of service, guaranteed and dynamic, help to define the four basic service models discussed in this book:

- The guaranteed-connections model
- The leased-line service model
- The dynamic-importance model
- The resource-sharing model

Another key issue when building a viable business model is *pricing*. From an ordinary customer's viewpoint, the main requirements for pricing are simplicity and fairness. But can simplicity and fairness really be reached in a multiple-service environment? If and when various parameters related to bandwidth, quality, and destination must be taken into account, pricing systems can become inherently complex and even incomprehensible. This chapter addresses various aspects related to pricing in multiple-service networks. A general model that takes into account bandwidth, quality, and availability is also introduced.

Before jumping into specifics, some general remarks on possible service models are in order. One question to consider is whether definite conditions are necessary concerning acceptable behavior. You may think that some terms, or conditions, are absolutely necessary. For instance, all TCP implementations shall be appropriate, or real-time service shall be requested only when needed. This definitely makes sense if the service model is based on the requirements of applications and on the cooperation of all end users.

On the other hand, if the service model is based on the price paid by the customer, the operator shall be more cautious when setting any conditions for the customers. There can surely be some restrictions, however; for instance, the provider might enforce conditions related to excessive or inappropriate use of email systems or other clearly undesirable behavior. From a traffic control viewpoint, however, it is up to the user to decide the actual purpose of the network service. If the customer sends more packets than his or her fair share, the network can simply discard some of the packets.

5.1 Service Level Agreement

Although the main aspects of a service level agreement (SLA) were introduced in earlier chapters, it might be helpful to review the basics. (The definition of SLA was presented in Chapter 3, "Differentiated Services Working Group.") An *SLA* is a contract between a customer and a service provider that specifies the forwarding service.

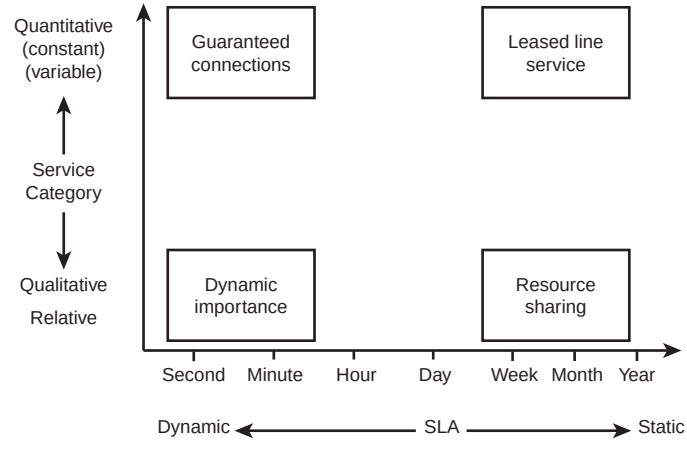
An important distinction between static and dynamic SLAs was also mentioned in Chapter 3. Static SLAs are based on a negotiation between human agents—that is, between customer and service provider; dynamic SLAs, on the other hand, change usually without human intervention and therefore require an automated agent (Bernet *et al.* 1998). A second distinction can be made among quantitative, qualitative, and relative service models. (See the section "Service Models" in Chapter 3 and the section 4.6.1, "Essence of Quality," in Chapter 4, "General Framework for Differentiated Services.")

As for the SLA between the customer and service provider, it is essential to define how the customer and network services are situated on these two scales, determined by dynamics and the service category. Figure 5.1 shows the following four primary approaches:

- Guaranteed connections (dynamic bandwidth)
- Leased-line service (permanent bandwidth)
- Dynamic importance (dynamic precedence)
- Resource sharing (permanent share)

Each primary term (such as guaranteed connections) illustrates the main application of the service models. The secondary terms (such as dynamic bandwidth) are more related to the technical implementation. The following sections discuss these models.

Figure 5.1 Four basic approaches for SLA.



5.1.1 Guaranteed Connections

The *guaranteed-connections model* is the traditional model of multiple-service networks, such as ATMs. Although this model often requires quite a lot of effort to implement, it is just as often considered to be the "right" target of service provision. Because some service providers will likely apply this approach despite the inherent difficulties, it is relevant to briefly describe how an SLA with guaranteed connections can be designed.

Note

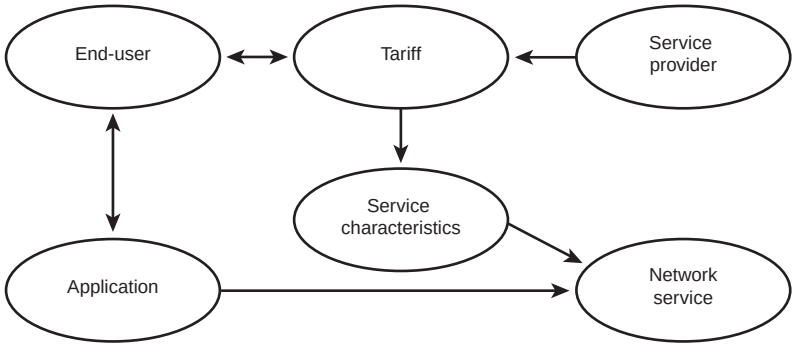
The term *guaranteed connection* refers here to a service used by an individual application with specific quality requirements and duration. In addition to this service model, there are other service models in which a customer buys a connection for aggregate traffic streams. In those cases, however, the connection is more permanent than what this section is assuming.

The feasibility of guaranteed connections was discussed earlier in this book: See section 2.4, "Integrated Service Model," in Chapter 2, "Traffic Management Before Differentiated Services," and the section, "Resource Reservation," in Chapter 4.

The SLA can be based on a model shown in Figure 5.2. The model is basically the same as the customer model presented in section 4.2.2, "Levels of Aggregation," in Chapter 4. The main actions of this model are as follows:

1. The customer decides to use an application.
2. The application informs the end user about required service characteristics.
3. The service provider offers the requested service at a certain price.
4. If the end user makes the decision to buy the service, the network provides a service based on the price paid by the customer rather than the requirements of the application.
5. The application begins to use the available network service as well as it can.

Figure 5.2 Service model for guaranteed connections.



It is assumed that the end user will select a network service that will meet the requirements of the application; however, there is no guarantee that this will happen. If a user reserves a 500kbps connection for an IP telephony call that actually requires only 20kbps, for instance, he may pay a lot of extra money without any notice from the network. (Although that is not directly a problem of the service provider, it may indirectly deteriorate the relationship between customer and operator.)

Although part of these actions can be automated, the basic process remains the same: The service provider sells network services based on individual connections, and the customer makes the decision based on the price and what the service offers. From the network service viewpoint, this means in essence that end users buy a fixed capacity more or less (usually more) from the network to a fixed destination, and applications then exploit that capacity as well as they can.

The main advantage of this model is that the customer is paying for definite service at definite price. It is a fair, clear, and consistent service model. The disadvantages of this model are the lack of scalability and its unfitness with adaptive applications. Consequently, this model is not realistic as the only end-to-end service model in IP networks.

5.1.2 *Leased-Line Service*

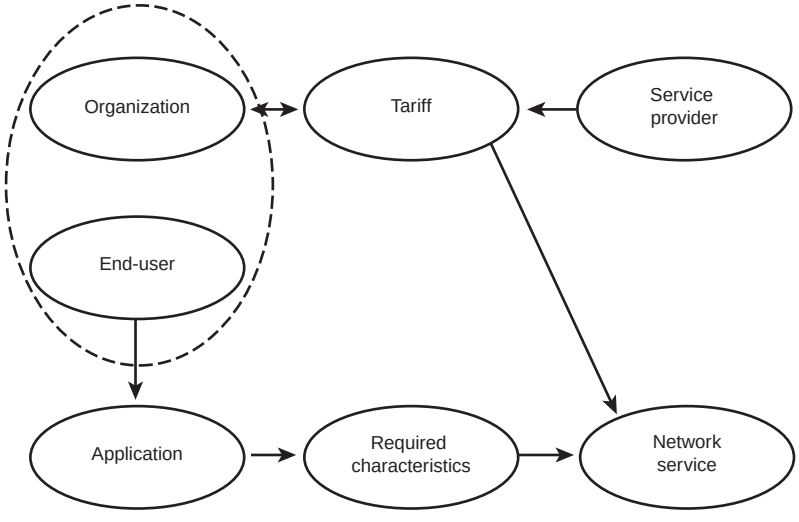
Unlike the dynamics of the guaranteed connection model in which the end user pays for a fixed amount of services for a fixed price, the leased-line service model is more permanent and less binding on the individual end user. Although this model can be used by and for individual end users, it is better used in an organization with a large number of end users—you are probably not eager to buy a permanent connection to all the places within the Internet you want to have a connection.

In this context, the main differences between these service models—leased-line and guaranteed connection—are the dynamic natures of the service(s) and the customer category. Leased lines are usually permanent, whereas guaranteed connections can be established within seconds. You can think of the guaranteed connection model as an abbreviation of switched guaranteed connections; leased-line service, on the other hand, represents all permanent guaranteed connections independent of the actual use of the service. When the term *virtual leased line* is used, it indicates that the level of guarantee is not as high as with "real" leased-line service.

Figure 5.3 presents one possible leased-line model. (See also section 4.2.2, "Levels of Aggregation," in Chapter 4 for an explanation of the differences among an application model, a customer model, and an organization model). An organization pays for a set of leased lines between its sites, and a number of end users then utilize the available network resources. Usually each end user starts an application, which then uses some network resources from the organization's common pool.

In a basic model, no congestion occurs inside the provider's network and, correspondingly, there is no need to define any mechanism for congestion. (Those mechanisms might be useful, however, in the private networks of the organization or at the interface between private and public networks.)

Figure 5.3 Service model for leased-line service.



There is a real demand for this type of service model. A large part of the traffic in Frame Relay networks is based on this type of service model. This model could even be relatively efficient provided that the load levels are stable enough to enable efficient network dimensioning. In practice, this could be possible if the number of active users is very large and the traffic process of one user is not very bursty. Based on the evaluation made in "the section A Model for Evaluating Statistical Multiplexing," in Chapter 4, it seems that it is not possible to satisfy this condition without controlling the traffic sent by end users.

Because this is not necessarily a realistic possibility, the use level of the leased-lines model remains low (or very low). This assumption is supported by some studies made on real networks. According to Andrew Odlyzko, "The greatest inefficiency in data networking today is that thousands of corporations are running their own private networks" (Odlyzko 1998). Whether this is an actual problem is an unclear issue because bandwidth is not necessarily the most expensive resource. Nevertheless, the situation provides an opportunity for more efficient use of resources.

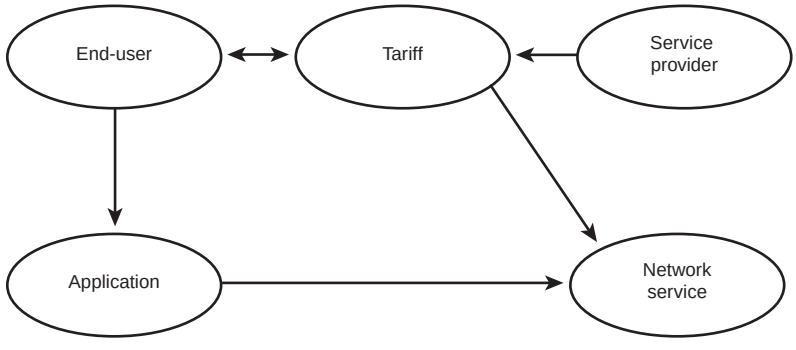
Although resource reservations can be considered to be against the fundamental principles of Differentiated Services, a couple of reasons make it possible to realize simple reservations inside Differentiated Services networks. The most obvious reason is that if the capacity reservations are permanent enough, they can be managed without any signaling system. The other, perhaps just as important reason, is that although the service model was based on the idea of reserving capacity, the traffic control inside the network does not necessarily rely on reservations for every individual connection but, for instance, on appropriate packet marking.

5.1.3 *Resource Sharing*

It is easy to proceed from the preceding leased-line model to the *resource-sharing model* by adopting the idea of using packet marking rather than real reservations. This model makes it possible to improve statistical multiplexing, but also necessitates the use of additional mechanisms to solve conflicts.

Figure 5.4 presents one practical service model for resource sharing. End users buy a share of network resources permanently. Applications exploit the available (variable) bandwidth as well as possible. In principle, there is no direct connection between application and network service besides the packets sent by the application.

Figure 5.4 Service model for resource sharing.



The main advantage of the resource-sharing model is that it provides a simple and consistent service: Each user gets the share that he pays for (provided that the underlying mechanisms can divide the network resources in a fair manner). Moreover, this model is inherently suitable for adaptive applications because there can be significant differences in the available share from time to time and from destination to destination.

This model (in its purest form) does not adapt well if the needs of the user and application change quickly and considerably. In particular, a real-time service can be necessary to make the overall service model attractive for a majority of users. Then an open issue is how to take real-time requests, or some other special requests, into account. If there is no incentive to request special service only when really needed, the result can be unfavorable for the service provider because everyone can ask for better treatment.

Another drawback to this model is the difficulty of verifying the performance and fairness of the service. There is no way for an individual user to verify whether he gets a fair share of the resources. It is, therefore, probable that the formal SLA is based on a somewhat different model from the actual service structure inside the network. The operator may, for instance, apply the qualitative service model within customer service, although the real implementation inside the network is based purely on a relative service model.

5.1.4 *Dynamic Importance*

The simple resource-sharing model described in the preceding section can provide acceptable basic service for most users and applications. Some needs cannot be satisfied with this model, however, because of its simplicity at the most basic level. Any one of the three major quality aspects—delay, importance, and bit rate—can be insufficient for a specific purpose. High-quality IP telephony needs better delay properties than those properties provided by the basic service. Some applications may need more bandwidth or higher assurance of packet delivery. Finally, some demanding applications, such as video meetings, may require all these characteristics at the same time. Some additional tools can be used to increase flexibility of static resource sharing, however.

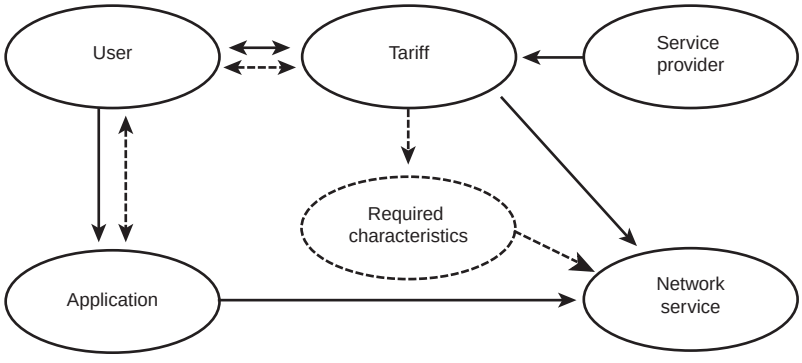
The key characteristic of this model (shown in Figure 5.5) is that a mechanism informs the network nodes that the current flow needs somehow better treatment than that which the user usually gets. These mechanisms are presented in the figure by broken lines.

Although this adjunct appears to be similar to that of the guaranteed-connections model, there is an essential difference: In the model of dynamic importance, you explicitly suppose that the network infrastructure is based on the resource-sharing model. Because of this, the dynamics should be based on changes in packet marking rather than reservations inside the network.

Note that this is in accordance with the basic philosophy of Differentiated Services. The most difficult issue in this system, provided that the basic resource-sharing system is available, is the price of extra quality. Basically, there are at least two options: a direct time price that depends on the required quality and a system price in which each user may utilize the monthly flat rate somewhat unevenly. If a user wants to momentarily send packets with

high bit rate, for instance, a smaller share during some other period can compensate for such a momentary "upgrade." (It is possible to apply a large variety of rules.) In the case of real-time service, an option is that if real-time service is requested, the share measured in bit rate is smaller than that of non-real–time service.

Figure 5.5 Service model for resource sharing with dynamic importance.



In all cases, the SLA should contain reasonable incentives for the user to request supplementary characteristics only when really needed. Moreover, the system should be as simple as possible, particularly if you suppose that the majority of traffic can be handled without these additional mechanisms. This seems to be the hardest part of this scheme: how to build an effective pricing system without too complex management and customer care.

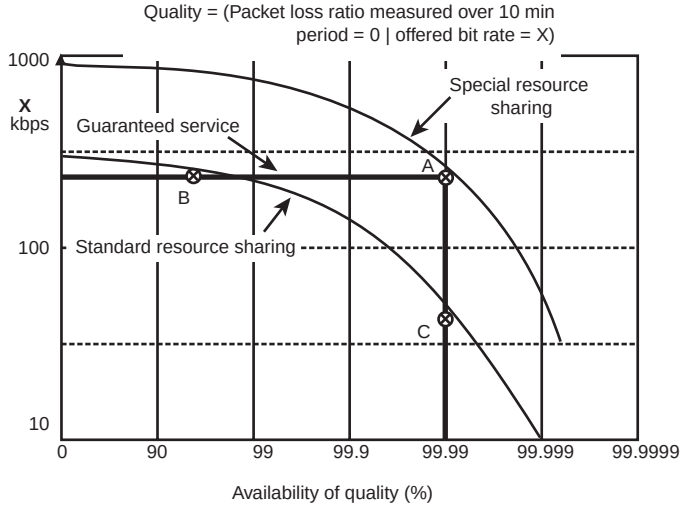
5.1.5 *Comparison Based on Availability of Quality*

The preceding section covered four service models that describe different kinds of SLAs. One of the fundamental questions of any quality of service offering is how the service provider can market it to customers. Two models—guaranteed connections and resource sharing—possess essentially different characteristics, whereas the customer's needs might be quite similar in both cases. What common aspect do both of these models share that makes it possible to market and compare both of these services? Availability of quality, which is discussed in more depth in Chapter 4, can be an appropriate tool for this purpose.

Suppose that a service provider has a customer who wants a 200kbps connection with high quality, for example. If the service model and the SLA are based on the guaranteed-connection model, the availability of quality can be depicted by the graph shown in Figure 5.6. The customer attains the desired characteristics by buying a 200kbps connection

through the network. A limit to availability exists, however, depending on the dimensioning of the network: With a finite probability, the connection request has to be rejected because of exceptional load level. Figure 5.6 illustrates the situation. Point A defines the availability of the quality in this case. For any bit rate higher than 200kbps, the availability is 0%; for any lower bit rate, the availability is 99.99% (or whatsever the call blocking is).

Figure 5.6 Quality function for guaranteed service and resource-sharing service.



The other option for the service provider is to build the customer service on the basis of shared resources. In this case, the service provider has a more complicated task when trying to define the availability of quality. Based on traffic and performance measurements and general experience, the operators of the network may be able to moderately distinguish the availability of different quality levels. Figure 5.6 shows an approximate availability function for a standard resource-sharing service.

Although this service may provide the 200kbps most of the time (point B in the figure), that may not be enough for demanding users and applications. On the other hand, the availability target can be met by a lower bit rate (point C in the figure). Still, that property is of minor value if the application definitely requires a connection of 200kbps.

The service provider has two main options to solve this dilemma. First, it can provide the possibility to buy a guaranteed service in addition to the resource sharing. Second, the service provider can merely provide different levels of resource-sharing services. It is even possible to base the service sold to the customer on the guaranteed-service model while the

implementation is done by means of a larger share inside the network. Figure 5.6 also shows a special resource-sharing service that meets this target.

5.2 Requesting Specific Service

The preceding section addressed issues relating to the contract between the customer and a service provider. Now the discussion takes one step toward technical matters, specifically how to request a certain service. Again, it is important to remember that the service can be based either on the resource-sharing model or the requested-quality model, and the dynamics of the request can vary from seconds to months.

5.2.1 Dynamic Quality or Bandwidth for Guaranteed Connections

The combining of dynamic allocation and guaranteed connections seems to be a difficult task. First, it requires some kind of signaling throughout the network. Because per-flow signaling inside the core network is beyond the scope of Differentiated Services, that possibility is not addressed further here. Notwithstanding, it is possible, as discussed earlier, that even though the customer makes a definite request by using RSVP, the request is converted to an appropriate Per-Hop Behavior (PHB) without any actual reservation for individual flows.

5.2.2 Permanent Bandwidth Reservation Versus Permanent Share

If the dynamic of reservation is days rather than minutes, the resource allocation task is essentially easier. Reservations can be made through the management system if the requirements do not change frequently. Of course, the management system and network nodes must be able to appropriately support the reservations.

From the requesting point of view, permanent share is similar to permanent bandwidth. The only fundamental requirement is that the management system and network nodes support the system; the issue of requesting a definite share seems to be simple.

The main difference may relate to the concept of *share*: One possible definition is merely to say that a customer obtains one share if he buys the basic service, and that all other shares are defined in proportion to this basic share. From a marketing perspective, however, this type of definition can be too abstract. For instance, a basic share may mean—under normal load conditions to most destinations—an available bandwidth of 50kbps with small packet-loss ratio.

More concretely, the service provider can use a description similar to that presented in Figure 5.6. The disadvantage of doing this is that the customer might consider the numbers

as guaranteed performance. If the service provider understands the situation in the same way, the number will be very low. For instance, the real available bandwidth frequently can be 10 times larger than the promised one.

Finally, the service provider can leave out all numbers and just state that certain service classes have lower and higher quality. Nevertheless, the realization of the service can be based on fixed shares. These three ways of requesting a service are possible; only hands-on experience can tell which is the best one.

5.2.3 Dynamic Share

With dynamic share, three different ways of messaging can be identified. (You should assume that the basic service model is resource sharing.) The customer may explicitly request a certain service, such as a 200kbps connection to somewhere using appropriate RSVP messages. In the Differentiated Services network, this message is translated into a proper PHB. In addition, the boundary node must have proper mechanisms to control the incoming packet flow, and probably a system to charge for the special service.

Another messaging possibility is when a customer does not request any definite quality characteristics but rather a bigger share, if the standard share appears to be insufficient. The advantage of this approach is that it does not necessarily yield any action outside the boundary node. Particularly, there is no need to change the PHB class, but only the thresholds that determine the importance levels within the PHB class. In this case, it is probable that an extra charge is needed to limit the requests of bigger shares. Various approaches do not include any additional prices—for instance, the monthly flat rate may include the right to use extra shares at certain times. Pricing is discussed later in this chapter, in the section "Pricing as a Tool for Controlling Traffic."

The third option is to use the DSCP field to inform the network about any special treatment needs. Packets belonging to an IP telephony service, for instance, can be marked as real-time packets with a suitable DSCP value. The boundary node then selects a proper PHB class for those packets. In addition, the system must have an incentive for the user to select real-time service only for those applications that really need that property. One incentive is to charge extra for real-time service. A more feasible solution, however, is to make the "real-time share" smaller than the default share while keeping the pricing the same. This latter solution is described later in section 5.3.6, "The Relationship Between Quality and Bandwidth."

5.2.4 Summary of Service Requests

These three alternatives for service requests are summarized in Table 5.1. Notice that although it is more convenient to use the term *user* (as the final entity making a decision),

in many cases the requests can be done automatically by the application without any human interaction.

Table 5.1 Alternatives for Service Requests

Way of Messaging	Content	Example	Boundary Changes Functions	PHB Change from PHB_default
RSVP message	Requesting quality parameters	Bandwidth = 200kbps	Traffic control and pricing	=> PHB_high
Share message	Requesting to change the share	Share_new = 2*share _old	Pricing and threshold	No change of PHB class
DSCP indication	Requesting special treatment	Low delay service	Pricing or threshold	=> PHB_rt

The main three quality aspects that can be requested are bandwidth, importance, and delay. An RSVP message is basically suitable for all of these. A customer can use an increased share either to increase bandwidth or to improve the importance level of the packets (that is, the availability of the service). These two changes can be presented in Figure 5.6 as horizontal or vertical shifts. A change of share does not necessarily have any effect on the delay characteristics, although that is possible.

DSCP indication is a particularly useful tool to inform the network about the need for real-time service, because most customers are not willing to choose permanently either real-time or data service. DSCP can also be used to indicate the relative importance of different packets, although it seems quite difficult to combine the resource-sharing model with the requirement to have different importance levels for different packets belonging to one flow. One reasonable scheme could be that packets marked as the lowest importance level have no effect on the PHB calculation mechanisms of any other packets; thus the user can effectively send a large amount of unimportant packets without deteriorating the treatment of other packets. Finally, DSCP indication is not the right mechanism to inform the network about bandwidth requirements.

Moreover, a network can use another alternative to prompt the user to select the right service. As discussed in section 6.2.3, “Feedback Information,” in Chapter 6, “Traffic Handling and Network Management,” networks can give information about the current load and quality inside the network. Because only a limited number of applications and

users can use the information, however, it is perhaps unreasonable to disseminate the information without an explicit request. Yet, it is possible to design a protocol by which the end users can inquire as to the current state of a path to a certain destination. The information could be related either to an individual PHB, a PHB class, or all PHBs.

Table 5.2 summarizes the characteristics of different approaches related to different applicability aspects. The additional aspects addressed here are scope of service, interdomain issues, and status of standards. Reservation mechanisms, such as RSVP, are suitable for connections with fixed endpoints—that is, for scoped services. (It is somewhat hard to imagine a real reservation without fixed endpoints.) The Differentiated Services–oriented approaches (share message and DSCP indication) are primarily appropriate with unscoped services; however, there is no technical obstacle to using DSCP indication with scoped services as well. Quality inquiry seems to require a fixed destination to be really useful. It may also be practical to certain a extent, however, even if the destination is not defined (just to get information about the general condition of the network).

Table 5.2 Applicability of Different Messaging Approaches

Way of Messaging	Quality Aspects	Scope of Service	Interdomain Flows	Status of Standards
RSVP message	Bandwidth, importance, delay	Scoped	Possible if available in all domains	Available
Share message	Bandwidth or importance	Unscoped	If the same service model is applied	Not available (useful, but not obligatory)
DSCP indication	Delay, importance	Unscoped or scoped	If possible to map PHBs	Standard PHBs can be used
Quality inquiry	Delay, packet-loss ratio (bandwidth)	Scoped (perhaps unscoped)	If support available	Not available, necessary

One general problem of any service provision within the Internet is that different domains may apply different service systems. RSVP could be used in some domains, for example, but probably never in all Internet domains. It should also be noted that if a user requests a certain service with RSVP, that does not mean that RSVP should actually be used in all domains (see the section "Interoperability with RSVP / Integrated Services" in the framework document [Bernet *et al.* 1998]). The advantage of RSVP is that complete standards define the formats of all necessary messages.

In a region with the resource-sharing model, standardization could be quite minimal. The main requirement of practical implementation covering a wide Internet region is that the resource-sharing model should be applied in some form by most of the service providers. The details of implementation can vary from operator to operator. Particularly, the communication between service provider and customer equipment can be based on proprietary messages, although some level of standardization would be useful.

5.3 *Pricing as a Tool for Controlling Traffic*

As mentioned several times in the previous sections, pricing is one of the key issues with any Differentiated Services model. This is an extremely complex issue and only some aspects can be addressed in this book. The main viewpoint in this section is Differentiated Services; for a more extensive discussion about Internet pricing, see "Internet Economics" (McKnight and Bailey 1997).

The main issue to be addressed is how a pricing scheme can help users maximize the ratio of user benefit to service cost. There are basically two options to tackle this issue:

- To somehow influence user behavior so that users do not waste network resources
- To give a fair service compared to the price paid by the customer (regardless of how the user is using the network service)

Within an organization, the first target is relevant and the solution is usually not based on actual pricing but rather on rules or recommendations concerning the use of network resources. Therefore, this chapter concentrates on the second case in which the price paid by the customers should somehow reflect the service provider's actual cost.

The framework addressed in this chapter consists of four main elements:

- Bandwidth
- Quality
- Availability of quality
- Price

If one factor is presented as a function of another factor, and the remaining two factors remain constant, six different cases result (as shown in Table 5.3).

Table 5.3 Basic Relations Among Main Elements Of Network Service

Case	Element	As a Function Of	For Constant
1	Price	Bandwidth	Available quality
2	Price	Quality	Available bandwidth
3	Price	Availability	Quality bandwidth
4	Availability	Bandwidth	Price quality
5	Availability	Quality	Price bandwidth
6	Quality	Bandwidth	Price availability

Cases 1 and 2 are the standard functions of pricing in reservation-oriented networks. The price of the connection is calculated as a function of bandwidth and quality, and the service provider usually tries to keep the availability constant—that is, the call-blocking probability should be approximately independent of bandwidth and quality requirements. There could be, nevertheless, a hidden availability aspect even in this case. When the price is higher for busy hours than for idle hours, it could be said that the price actually depends on the availability.

To get a relevant insight into all these relations, a simple but somewhat realistic mathematical model may be helpful. The following exemplifying figures are based on one formula of the form, as shown in Formula 5.1:

Formula 5.1

$$\log(P) = c_B*\log(B) + c_Q*\log(Q) + c_A*\log(1-A) + c_P$$

In Formula 5.1, P = price, B = bandwidth, A = availability, and Q = quality (delay variation in the following examples, but it can be another quality parameter as well). Formula 5.1 is selected here mainly because of simplicity. The logarithmic form makes the effect of each parameter systematic. If bandwidth is increased by a factor of 2, and c_B is 0.5, for example, the price is increased always by 41% regardless of the other parameters.

It should be strongly emphasized that this is just one example of the many possible pricing schemes, and particularly that the selected constants for c_B, c_Q, $c_{A,}$, and c_P are arbitrary (although they try to be realistic). Further, the position of this book is that the basis of pricing should be as consistent as possible, even though the actual tariffs may deviate from any simple mathematical formula. (There are many opportunities for inconsistent tariff structures.)

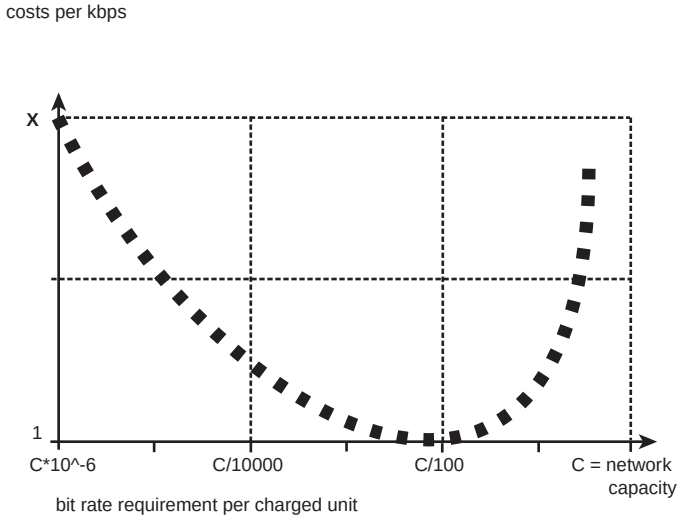
5.3.1 Price of Bandwidth

Figure 5.7 depicts one fundamental problem associated with the pricing for network services related to the relationship between bandwidth and price. On the one hand, practical experience shows that when bandwidths differ significantly, a linear relationship is not a practical approach. According to "Dynamic Behavior of Differential Pricing and Quality of Service Options for the Internet," for example, if the price of a 56kbps connection were \$595, the price for a 1.5Mbps connection would be \$1,795, and the price for a 45Mbps connection would be \$54,000 (Fishburn and Odlyzko 1998). In this case, the relationship between bandwidth and price is far from linear.

These figures strongly support a model in which parameter c_B is significantly smaller than 1. This phenomenon can be explained by the following example. Assume that the average bit rate required by a customer is R_{ave}, and the total cost of providing service for all customers is $C_T(R_{ave})$. If the average bandwidth requirement is increased tenfold in a way that all other factors are kept unchanged (as far as it is possible), the total cost of service provision is evidently more than $C_T(R_{ave})$, but most probably significantly less than $10*C_T(R_{ave})$.

On the other hand, if you consider the dimensioning problem of a given network, you might come to a totally different conclusion. One connection with a constant bit rate of 1Mbps consumes basically as much resources as 10 connections with a constant bit rate of 100kbps. If a very large number of tiny connections are made, however, the management costs could induce the major costs. Therefore, the main issue may be the charged unit measured in bit rate rather than the average bit rate of a connection. Consequently, for a given network the result can be something like that shown in Figure 5.7.

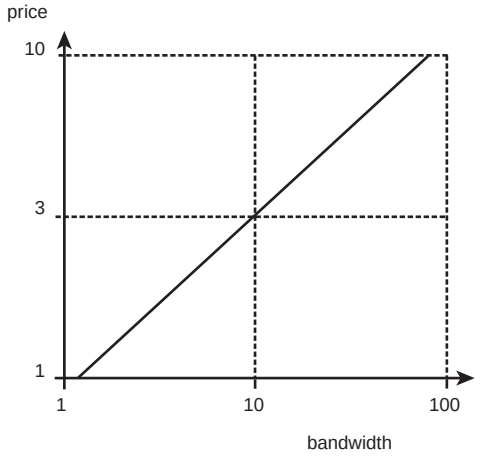
Figure 5.7 An approximate relationship between charged unit and costs.



So there are two opposing phenomena: The total cost tends to be high if the charged unit is very small, and the effects of statistical multiplexing tend to deteriorate if the number of independent units is small (perhaps less than 20). In the middle region, the cost per bit rate could be relatively constant. It should be stressed that the right ascent in the figure is relevant only if you suppose that the total network capacity is fixed and cannot be easily updated and that the operator's intention is to use the network for public network services.

In reality the situation is not static, but highly dynamic. Network capacity is updated all the time according to demand, and in high-capacity public networks one customer is seldom so dominant that the effect of statistical multiplexing considerably deteriorates. Therefore, despite the significant opposing arguments, it is possible to tentatively apply Formula 5.1 with a C_B smaller than 1. A value of 0.6 is used in Figure 5.8, as well as in other figures related to pricing. Service providers may, of course, build real services based on totally different approaches.

Figure 5.8 A tentative relationship between bandwidth and price.

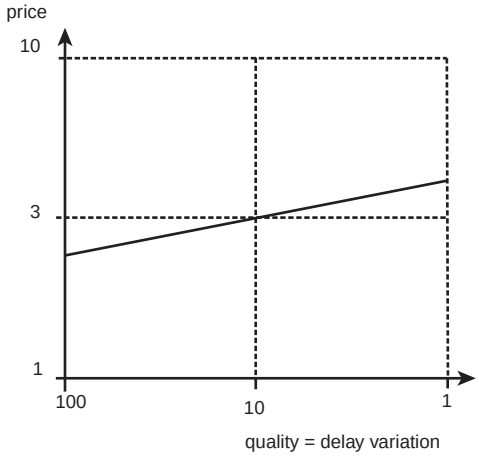


5.3.2 *Price of Quality*

Quality is an ambiguous term both generally and particularly from a pricing point of view. This attempted concise evaluation, however, concentrates on one quality aspect: delay variation. (Note that bandwidth aspects are already discussed, and the next topic, availability, contains aspects related to different packet-loss ratios.) Because real-time network service requires additional control mechanisms as well as additional management actions, it is justified to suppose that the price of real-time service should be somewhat higher if other aspects remain constant. Figure 5.9, for example, supposes that a 100-fold decrease in delay variation means a double price.

The main point seems to be that real-time support makes traffic control more complicated. However, it is difficult—even practically impossible—to give any clear rule for the price difference. In some cases, for instance, when the traffic flow sent by the user is exactly constant, it is not clear whether there is any significant difference from a statistical-multiplexing or traffic-control viewpoint as to whether a flow uses real-time or data service.

Figure 5.9 A tentative relationship between delay variation and price.



5.3.3 *Price of Availability*

The relationship between availability and price may appear somewhat artificial (if you suppose that the availability is a common characteristic for the whole network). One possible interpretation is that availability is related to the availability of a service at a certain price at different times. When availability is high (say, 99.999%), the service is available even during the busiest times of the year. An intermediate availability means that the service is available at that price on a typical busy hour. Finally, low availability means that the price is valid on idle hours. A value of 0.15 is used for constant c_A in Figure 5.10.

A similar model can be used if availability is interpreted as the probability that a packet is successfully forwarded through the network. For this, however, you must suppose that availability means primarily the probability that all packets are forwarded during a moderately long period for reasons discussed in section 4.2.1, “Availability of Quality,” in Chapter 4.

Figure 5.10 A tentative relationship between availability and price.

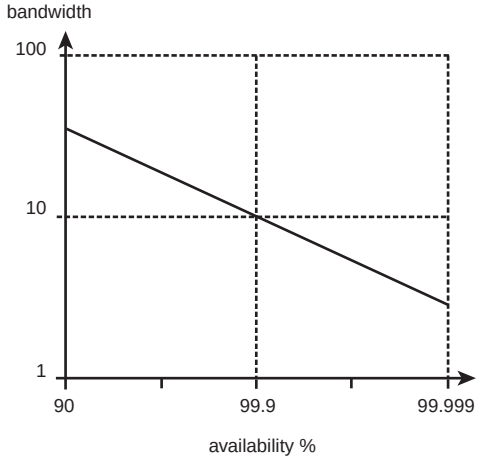
5.3.4 *Relationship Between Bandwidth and Availability*

All the previous evaluations were based on the assumption that price is a variable quantity. This is a reasonable assumption if the operator applies service models with reserved connections. On the contrary, with some other service models it is not realistic to expect that a different price can be attached to all differing situations. Particularly if the service provider uses flat rate pricing, the price is basically fixed whereas all other parameters can change.

Figure 5.11 shows the relationship between bandwidth and availability for fixed price and quality, for example. It is easy to draw the figure, because the same parameters as in the earlier sections can be used. Yet, the reality is much more complex because with many service models the relationships between different aspects, such as availability and bandwidth, are results of a complicated process that is not totally controllable by the service provider. Therefore, all the figures are definitely illustrative, but may be used to check whether the parameters chosen earlier are realistic.

Essentially, the relationship between bandwidth and availability depicts the differing demands on busy and idle hours (and minutes). If the target is to share the bandwidth equally on times with differing busyness, the availability-bandwidth relationship directly reflects the relationship between demand and time of day. It is apparent that there is a significant difference between available bandwidth during busy and idle moments, perhaps of the order of 10 as in the tentative model. This difference, itself, encourages end users to use less bandwidth during busy hours, or to change the time of use from busy to idle times if the application is not adaptive.

Figure 5.11 Bandwidth as a function of availability for fixed price and quality.



5.3.5 *Relationship Between Quality and Availability*

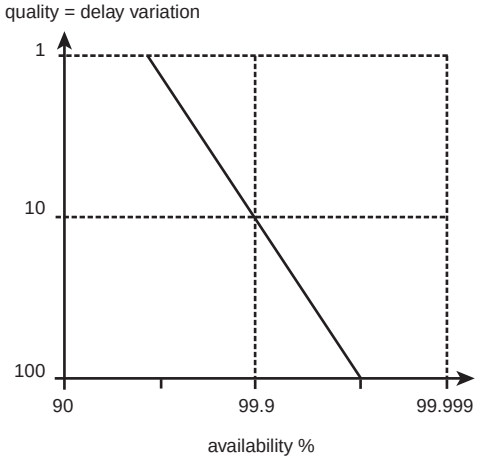
If the application cannot adapt to available bandwidth, it may be possible to degrade the quality in case of insufficient bandwidth. If we again apply Formula 5.1, we get the relationship between quality and availability shown in Figure 5.12. It seems that there is usually only a limited possibility to apply this approach. If, for instance, a real time application requires definitely small delay variation, it is probably not reasonable to allow much larger delay variation during the busiest hours.

Consequently, the main effect of this relationship could be that there is an incentive for customers not to use high quality service for less demanding applications in particular during busy hours, that is, when network resources are scarce.

5.3.6 *The Relationship Between Quality and Bandwidth*

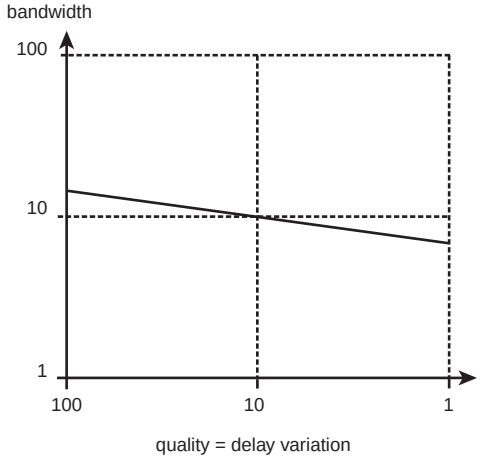
Finally, it is possible to assess the relationship between quality and bandwidth. It seems that this relationship is relatively weak in the sense that even a substantial change in quality may cause a relatively small change in available bandwidth. (However, that statement is valid only if the general model is valid, which is not at all sure.) Nevertheless, it is important that users can attain more bandwidth if the quality requirements are looser, if it is not possible to attach additional pricing for higher quality.

Figure 5.12 Quality as a function of availability for fixed price and bandwidth.



The available bandwidth is approximately three times larger for non–real-time service than for real-time service in the model shown in Figure 5.13. One interpretation of this tentative result is that the average load level of real-time service classes can be one third of the load level of non–real-time services in cases where the whole traffic uses only one service class. Note again that this specific relationship between available bandwidth and delay characteristic is a result of selected parameters in Formula 5.1, whereas in reality the relationship could be totally different.

Figure 5.13 Bandwidth as a function of quality for fixed price and availability.



5.3.7 *Effect of Variable Destinations*

One important aspect that was not addressed in the tentative pricing model is the effect of destinations. The Internet is obviously a very heterogeneous network. Some parts of the network are equipped with high-capacity links and routers that can handle all the incoming flows without any losses most of the time. Then there are low-capacity access networks with a permanent lack of resources. Finally, some links, most prominently those between main continents, are expensive and heavily used.

If the resource-sharing model described in the beginning of this chapter is applied, the real available share depends strongly on the destination. The fair share for an ordinary customer on a local link may be 200kbps. The fair share over an Atlantic link may well be fraction of that, however, say 20kbps.

A fundamental consequence is that either the pricing will depend somehow on the destination or, alternatively, a higher importance level is needed to transmit packets over expensive and highly loaded links. In any case, the customer has to pay for the availability of the service, where availability is related not only to point of time but to destination as well. Therefore, the model presented in Figure 5.11 might also be applicable to this case. By decreasing the bit rate enough, the customer obtains the "right" to use even the most expensive links. From the network-control perspective, that means that those packets must get a high-importance marking.

5.3.8 *Levels of Pricing*

There are two basic levels of pricing. In residential markets, each individual customer is paying for his or her service. In the case of organizations, however, the whole service is usually paid for by the organization. Then there are sometimes additional needs related to the intermediate levels of the organization. Even though the contract between an organization and the service provider can be based on the total service, there is often a need for internal pricing of departments, based on resources they have used. The organization can accomplish this in one of three ways:

- The simplest approach is just to collect information about the use of network resources and make internal charges based on the information. This system has to take into account the requested quality levels, for instance, by measuring separately the load on each PHB in a Differentiated Services network.
- In a more sophisticated system, the whole organization may have a common resource pool, departments may buy a share of that pool and, finally, each end user may have a share of the department's pool. Although this appears to be a desirable system, it may increase considerably the complexity of traffic control and the burden on management personnel.

- In an intermediate approach, the management system has a large pool for the whole organization, and departments can reserve or buy a definite share for each end user. When the total pool is increased, each customer immediately recognizes a similar increase of available capacity.

5.3.9 *Variable Bit Rate*

This chapter assumes that the bit rate of each flow is more or less constant. That is, of course, an unrealistic assumption because Internet traffic is a highly variable entity. Many recognized studies relate to the optimal pricing of variable bit-rate connections (Roberts, Moggi, and Virtamo 1996).

An elegant approach, proposed by Frank Kelly, is based on the concept of effective bandwidth (Kelly 1996, 141–168). In Kelly's method, pricing of a connection depends on the following three parameters:

- Peak rate
- Mean rate declared by the customer
- The real (measured) mean rate

The better the customer is able to predict the real mean rate, the lower price she gets. An example of this is in cases where a service provider offers reserved connections for a price depending on the quality and bandwidth requirements. This method is, therefore, difficult to apply in Differentiated Services networks. (Yet the fundamental idea that the user gets a certain advantage if she can accurately predict her bit-rate requirement can be useful if a network operator wants to use network resources very efficiently and all the required mechanisms are available.)

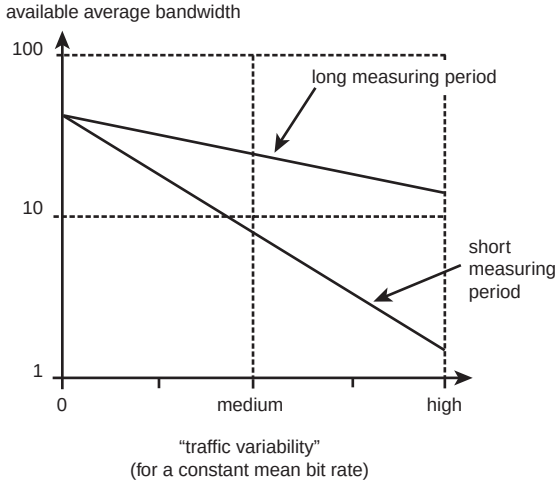
What can the service provider do in the case of the resource-sharing model? A direct relationship between traffic variations (or the user's ability to predict mean rate) and pricing seems to be impractical. In contrast, the available average bandwidth may depend on the amount of traffic variations. This kind of relationship can actually be relatively easy to realize. If the resource sharing is done purely based on the bit rate measured over a short period, constant bit-rate flows may attain significant advantage over variable bit-rate flows.

This relationship can be, to some extent, adjusted by changing the measuring period. A short measuring period strongly favors constant bit-rate sources, whereas a long measuring period results in a more equal share between constant and variable flows as shown in Figure 5.14. Note also that a very long measuring period yields the same result as usage-based pricing that is insensitive to traffic variations.

The optimal length of a measuring period depends both on the service model and on the requirements of traffic control. A short period could provide better means for effective traffic control, if you suppose that the main task of traffic control is to alleviate congestion. From that perspective, it seems reasonable to penalize those flows that have the most significant effect on congestion—that is, flows with highest momentary bit rate. This is particularly important with real-time services with small buffers that can be filled with short, but intense, traffic bursts.

On the contrary, because non–real-time services with larger buffers can better tolerate traffic bursts of short duration, the measuring period could be longer for those services. Consequently, the two lines in Figure 5.14 can also illustrate the pricing difference between real-time and non–real-time services.

Figure 5.14 A tentative relationship between traffic variability and available bandwidth.



Summary

This chapter discussed primarily two main aspects related to the customer contract, or service level agreement: the service model and the pricing model. Four service models were identified in the first section of this chapter. The models differ both in the dynamics and in the level of guarantees:

- *Guaranteed connections*: This is a general term for a service model in which a customer can request a connection with specific quality requirements. The service provider either offers a connection with the required characteristics or, if there are not enough resources, rejects the connection request.

- *Leased-line service*: This is a general term for a service model in which a customer, usually a large organization, buys a permanent connection with a constant bit rate through the network. The quality, including security aspects, should always be high enough for critical business needs.
- *Resource sharing*: This is a general term for a service model in which the customer buys a share of network resources instead of specifying the requirements of individual flows. The actual amount of resources the user can obtain depends inherently on the network's load level. In this model, the assumption is that the size of the share is relatively permanent.
- *Dynamic importance*: This is a general term for a resource-sharing model with improved dynamics. In this model, each user is allowed to request dynamically higher importance classification, either for individual packets or for all packets during a short period of time.

The main conclusion related to pricing is that *the overall pricing model has to be very consistent*. One specific pricing model was introduced, mainly for illustration purpose. The main property of this tentative model is that it defines the relationship between any two aspects if all other aspects are kept constant. If a customer is paying a constant bill every month, for example, it is still possible to have quality differentiation if the bit rate is changed at the same time.

Nonetheless, the real implementation of service makes it often impossible to realize a given pricing model. Therefore, it is necessary to design the service and pricing models jointly with the development of traffic-handling mechanisms used inside the network.