CHAPTER 1

The Target of Differentiated Services

Service differentiation is an old topic in many business areas. Nevertheless, if you search for references to Differentiated Services or service differentiation dated 1996 or earlier, you probably won't find much about the Internet and its services. During the past year, however, Differentiated Services of the Internet has become a popular concept. But why do we need a novel concept within the specific field of the Internet? As Ludwig Wittgenstein (1889–1951) said, "He who controls vocabulary controls thought." In the best case, a new concept with a new vocabulary provides a useful framework for the development of new ideas and the analysis of old ones.

This chapter explains the main motivation behind the Differentiated Services effort: how the Internet has changed and how this change has altered the requirements for Internet services and traffic management. I introduce the fundamental building blocks needed to realize reasonable customer service, and the desirable characteristics at all levels of implementation. These characteristics, or attributes, are then used as a thread throughout the whole book. This chapter elucidates the philosophical basis of the book.

Note

Although it could be useful to give an exact definition of Internet service, that is a somewhat risky approach because the Internet service model is still evolving and is prone to significant changes. Chapter 4, "General Framework for Differentiated Services," further illustrates the various aspects of the service models. Which one of them will be prevalent in the future is still an open issue—and the development of Differentiated Services may have a crucial role in that service evolution.

1.1 The Core of Differentiated Services

Imagine yourself as a mechanism inside the Internet, something like an intelligent sorter in a post office. You are working at a service provider that transmits packets between end users. The service provider in this case has different end-to-end services using different "postage." More expensive services may require quicker treatment inside the "post office" as well as different transmission tools, such as a "courier" rather than a "mailman." The network needed to accomplish this task consists of a number of packet-handling centers and paths between them.

When you, working in one of those centers, receive a packet with certain information in its tag (or *header* in Internet terminology), you will decide how to treat the packet. You can choose from only a limited number of different actions:

- You can deliver the packet immediately, before all other packets, because you consider it very urgent.
- You have a number of boxes in which you can put the packet waiting for delivery.
- You can totally discard the packet because you think that you cannot deliver all incoming packets anyway.

How has the Internet changed from the viewpoint of this mechanism? What changes are coming in the near future?

Previously, your task would have been relatively easy. Basically you would have needed to look at the address on the packet and then, based on a routing table, you would have forwarded the packet in the right direction. From time to time, some packets marked as *urgent* would have arrived, and you would have had to deliver them as soon as possible. If the box (or queue) seemed to overflow, you would have had to discard some incoming packets to lighten your load. You could have expected that if you had to discard some packets, the senders would have been informed about the situation and, consequently, they would have sent packets more slowly. Finally, if there had steadily been too many packets for your capacity, you would have had to be retired and a new handler with a tenfold capacity would have been substituted for you.

In this case, most of the packets deserved equal treatment, independent of the sender or receiver; there was not much need for calculating how many packets someone had sent during the past hour or what kind of packets were waiting for delivery.

And everything worked fine with this simple scheme. Why? Mainly because the Internet population was relatively coherent in the sense that the communication between Internet users made it possible to build the Internet based on the principle of reciprocity.

Reciprocity means that anyone is allowed to send a large amount of important packets because other users could expect the following:

- They could do the same when necessary.
- No one wasted network resources by sending useless packets.
- Most users decreased their sending rate when capacity limits were exceeded.

The Internet community, with both network engineers and users, was coherent enough for this kind of system 10 years ago. Things have changed during the past few years, however, and you cannot—whether in your actual role as a network device or not—expect that all packets are equal anymore. In the same Internet, fundamentally different packets are delivered: Some packets are vital to someone's business; some packets become obsolete within fractions of a second; a lot of packets are sent just for curiosity; and some packets contain information that could be considered totally valueless by some other users.

Now you, as a network device, must decide how every individual packet will be treated. The core of the problem is that the current Internet environment is heavily fragmented. No single cooperative group sends packets; instead, a large number of groups and even separate end users—all with potentially different desires, different requirements, and a different willingness to pay for different properties—are sending packets. To treat all the packets fairly seems to be almost an impossible task, even if you have all the necessary information and enough time to make reasoned decisions.

This is the very area of problems that Differentiated Services aims to resolve, this "fragmentation" problem. This book seeks to clarify exactly what Differentiated Services is and how you can maximize its utility. The main goal of this book is to present a consistent view of the development of the Internet toward Differentiated Services, using a limited number of key concepts. The introductory concepts discussed here relate to the components that either are crucial for building the appropriate Differentiated Services or have a significant effect on the whole system of Differentiated Services.

1.1.1 Basic Entities of Differentiated Services

Differentiated Services consists of an array of technical issues, but that's not all. Differentiated Services must also be understood as having an inherent business and even psychological aspect. To provide a common reference point for this discussion, it is important to first identify the basic entities of Differentiated Services. Three of these basic entities

were introduced in the preceding example of a packet-handling center and are explained in this section:

- The service provider
- The end user
- The mechanism that treats packets in different ways

After briefly discussing these basic entities, this section introduces three other important entities of Differentiated Services: applications, networks, and vendors.

The Service Provider

The term *service provider* is often used to refer to two different things: the actual service provider responsible for customer relations (the broad sense of the term, and the way in which it is used here), and the network operator responsible for operation and management of the network. Although not used in the first simplified illustration of Differentiated Services, this distinction is necessary when assessing several issues, such as business models or interworking. (See Chapter 8, "Interworking Issues," for more information.)

In an extreme case, a service provider can be just a brand name (much like Coca-Cola or Nokia). All the technical devices required to implement the marketed service are collected in the same way that many branded articles are produced and marketed all over the world—that is, without having much more in common than the name. Subcontractors make the actual product (or service), even though the end users might think that they are doing business directly with the holder of the brand name.

At the other extreme, a service provider may be responsible for all parts of the service—from network construction to customer care. In addition, the relationship between end users and the service provider can vary significantly: The service provider might work within the same corporation as all the end users, or the service provider might sell small pieces of service(s) to a large number of individual users. An enterprise that brokers bandwidth and services between users and multiple providers is another alternative. Yet a more complex situation is when several brokers negotiate with each other. As this discussion hopefully makes clear, the term *service provider* applies to many different business models (both impossible and unnecessary to elucidate exhaustively here).

End User

An *end user* is you or any other person who is using the Internet for any purpose. The key point is that the end user is a human being, even though the chain from the real bit transfer to final end user can be long. If the network is used to convey data automatically from a meteorological station to a host computer, for example, some human need is still

prompting that bit transfer request. It is fair to suppose, therefore, that end users have a variety of emotions that influence the use of the service.

If you think that you are getting poor service, for example, you can change your service provider (regardless of what is actually causing you to think that the service is poor). From the service provider's point of view, your reasoning might actually be irrelevant, flawed, irrational, and / or based on limited information. The applications running on your own computer might be incorrectly configured, for example, or you might have heard that your provider is somehow unreliable. The important thing to remember is that customers do not always make their selections based on technical facts.

Mechanism

The third basic entity, the mechanism used to illustrate the principles of the network, is an integral part of Differentiated Services. Technically speaking, it can be called a *mechanism*. A mechanism can be understood here to be any piece of equipment or software that does a particular job inside the network nodes. A typical example is a device that categorizes packets into two classes of importance based on some rules defined in the service-level agreement.

Applications, Networks, and Vendors

End users use *applications* to satisfy some demand that can vary from serious to entertaining. The demand can be to find product information, to converse with a colleague abroad, or to spend some time surfing the Net. At this point of the discussion, the particular use or aim of the application is not important; it is important, however, to notice that the underlying need is rarely just to transfer some bits through the network.

The term *application* should also be understood broadly: It covers typical user applications as well as all protocols not controllable by the service provider, such as TCP / IP protocols running in customers' computers. In contrast, the *network* as a basic entity is something totally managed by the service provider and used to transmit information from one end user to another.

Finally, *vendors* supply network components (both hardware and software) to service providers, network operators, and end users. Without these components, Differentiated Services would be an empty idea.

Note

This introductory picture should be refined. One principal circumstance not yet addressed is that several service providers that are connected to each other may provide the same services (more or less). Moreover, a Differentiated Services network does not form an insulated region; instead, a lot of other network technologies are used in parallel with it—and all these networks raise interworking concerns. The success of Differentiated Services depends crucially on how effectively these concerns are addressed. This is a substantial topic and is covered more fully in Chapter 8, "Interworking Issues."

1.1.2 The Relationships Among the Basic Entities

Something crucial is still missing, however: Although the preceding section drew a skeletal outline of the basic entities, it is important to flesh out that skeleton by defining the relationships among the basic entities. To accomplish anything useful on the Internet, an awareness and understanding of these relationships is necessary.

The relationship between end user and service provider can be called *customer service*. Customer service includes all the issues that have a significant effect on customer satisfaction. The issues encompass both technical details, such as packet-loss ratio, and nontechnical details, such as the friendliness of help desk staff. The formal part of customer service can be called the *service-level agreement*. The service-level agreement may relate to a specific need of transmission, such as connection to a video server, or it may specify some general issue, such as the appropriate response time for customer technical support. In addition to the formal service-level agreement, customer expectations play an important role when customers are assessing the quality of a product (in this case a service), even though these expectations are not recorded in any document.

Another fundamental relationship between user applications and the network can be called *network service*. There are a lot of network services that vary in characteristics, such as packet-loss ratio, delay variation, and available bit rate. In other words, network service defines what an application is supposed to do on a technical level, in such a way that the quality level is usually straightforward to measure.

The relationship between the network, which often covers all possible technical issues, and the mechanism, which is apparently part of the network technology, tends to be artificial. If you think of the network as a unit that has a general purpose, however, you can use the term *traffic handling* to refer to this relationship; this term illustrates the low level of service provided by mechanisms.

The relationship between the service provider and the network and mechanisms can be called *operation and management (OAM)*. OAM covers such issues as building reasonable services based on available mechanisms, making traffic measurements for network-planning purposes, and solving fault situations. In many cases, OAM is the main single cost factor—in particular, if any of the OAM functions require manual operations (and usually they do).

Finally, the vendors provide applications and network components. Although vendors are definitely crucial to applications, services, and networks, they do not usually have any specific role in the actual service. Therefore, it is unnecessary to define any particular relationship between vendors and other basic entities. It is necessary from time to time, however, to remember that there are extra players, called vendors, in the field and that these extra players have their own specific interests.

1.2 The Four Attributes of Differentiated Services

The many special ingredients of Differentiated Services have yet to be described *per se* here. To that end, it is important to regard the *differentiated spirit* as the principal target of the whole effort and to consider the four attributes discussed in this section as secondary targets.

End-user needs should be the paramount focus of any Differentiated Services approach. Unfortunately, most end users have no idea what their future needs will be, particularly with regard to Differentiated Services. This discussion, therefore, reviews the following four universal attributes of Differentiated Services:

- Fairness
- Robustness
- Versatility
- Cost efficiency

These attributes often intertwine, and a Differentiated Services approach might also apply other attributes to a job at hand. These four specifically listed attributes relate to this discussion, however, because they can be applied both at the customer-service level and at technical levels; in addition, together they can cover all the key aspects discussed earlier (provided the terms are used generally).

1.2.1 Fairness

The concept of fairness relates directly to the essence of human viewpoint. For business managers and administrators, it is of great importance to thoroughly comprehend what customers want and what they think or "feel" about the service. It is not enough to look at money and technology only. The other three attributes focus on the hard values, such as earnings, efficiency, and reliability. *Cost efficiency* emphasizes the need to assess technologies from a realistic business viewpoint, for example, and *robustness* calls for reliability.

One of the several meanings of *fair* is concisely expressed as "reasonable according to most people's ideas of justice" (taken from *Longman Language Activator*, Longman Group UK Limited, 1993). This definition emphasizes the emotional aspect of the term. Note also that *fair* does not necessarily mean equal treatment; fair just means that the treatment is acceptable to *most* people. In addition to the psychological aspects, technical aspects of fairness must be considered when appraising the properties of mechanisms and network services. These aspects of fairness are secondary, however, in the sense that fairness is finally assessed in a person's mind and is not very well represented by any mathematical formula.

Fairness is the key attribute of the relationship between the end user and the service provider. Whatever the service provider is selling to customers, it must be regarded—first and foremost—as fair by customers. In fact, the fairness of the service from the customer viewpoint is the first issue to be addressed when this discussion turns to other networking technologies in Chapter 2, "Traffic Management Before Differentiated Services," and Differentiated Services proposals in Chapter 7, "Per-Hop Behavior Groups."

Fairness Versus Quality of Service

You may be curious about why the popular term *quality of service (QoS)* has not been introduced into this discussion. The reason is that the QoS concept is often used in a limited sense, in which it means support for service with certain predefined characteristics that can be directly measured. In the case of communication services, however, typical measures of service are maximum delay and loss ratio. With regard to these measures, a premature or excessive use of technical parameters can lead to somewhat misleading conclusions. In particular, it should be noted that technical parameters cannot cover all the substantial aspects of customer service—most users are hardly willing to consider technical details, and even fewer users will make measurements to verify the actual quality of service.

To express it simply, most customers just assume that the market somehow establishes the right price level (whatever the word *right* means). An average customer just decides whether the price offered is low enough to justify buying the product. Nevertheless, the customer wants to be sure that the market is fair—that is, that everyone pays the same price for the same product. This is the essence of fairness.

Because the essential characteristics of customer service can be more precisely discussed using the word *fairness* rather than *quality*, *fairness* is used as the key term. In this book, the term *quality of service* is used only when it is possible to accurately define the required characteristics and to verify whether the service actually meets these requirements.

Groups and Fairness

Before diverting this discussion to technical matters, it is important to understand another very important concept: the *group*. A group is a set of entities located close together or classed together. The basis of classification is often a certain quality that each individual entity has in common with the other entities so classed. Why is *group* such an important term? Because it is impossible to evaluate fairness without having a good understanding of the group to which an end user or a packet belongs.

A group must be coherent to be meaningful. (*Coherent* here refers to something that has unity of ideas and / or interests.) Within a coherent group, each member is somehow

responsible for the behavior of all other members, or, at least, this is considered fair from an outsider's view. The other side of the same issue is that each member of the group expects to benefit from the group membership.

An example that relates to Differentiated Services might help to explain the use of the word *group* in this discussion. A similar contract between several end users and one service provider makes a basis for an inherent group. The source of cohesion is the contract between the user and the service provider. The user joins a group with certain rights and responsibilities, partly described in a service-level agreement and partly based on common sense. Each user expects certain predictable behavior from all users within the group, and will be content with the service. In a best-effort service, for example, a user buys access to a network. That access has a certain physical bit rate. The user won't (or at least shouldn't) expect to obtain any definite bit rate from the network; instead, the user should expect a fair share of bandwidth and be content with that service.

Some service-level agreements, or contracts, might differ slightly from one user to the next. These relatively minor differences do not necessarily justify the formation of several groups. If that were to happen, the management of the total system would quickly become too troublesome. On the other hand, it is difficult for end users to assess the fairness and other key properties of a service if contracts vary too much—that is, there are a large number of small differences from contract to contract. One reasonable solution to this problem is for the service provider to offer only a couple of service levels. (The airline industry is using this approach to market and manage their services.)

Fairness and Service Provision

At all times, potential for group to overlap exists, possibly with a certain hierarchy. Unfortunately, it is not always clear which grouping is relevant in each case. To further illustrate this complex issue, consider the following example, "Limiting the Load Level to Avoid an Overload."

Note

This book uses a fictitious company and service to provide concrete examples of various aspects of implementing Differentiated Services: Fairprofit, an Internet service provider, and Quicksure, a service supplied by Fairprofit and other ISPs that provides reliable service for real-time applications.

Limiting the Load Level to Avoid an Overload

Several Internet service providers, Fairprofit among them, share a backbone network for transmitting traffic between customers of different service providers. All providers have the same service structure, including Quicksure. Now the combined load of Quicksure from Fairprofit customers significantly exceeds its normal traffic level, even though every individual user complies with his or her service-level agreement. In consequence, the backbone operator is compelled to somehow limit the load level because of the imminent overload situation. The operator can apply several different approaches to manage the situation:

- Limit the traffic of the customers belonging to both Fairprofit and Quicksure groups.
- Limit the traffic of all Fairprofit customers independently of the service group.
- Limit the traffic of all Quicksure users independently of the service provider.
- Limit the traffic of all customers belonging to either Fairprofit or Quicksure.
- Limit the traffic of all customers.

Further, these approaches can be combined in different manners—for instance, by more tightly limiting all traffic from Fairprofit and less tightly from traffic belonging to the Quicksure service (independently of the service provider). Now the fundamental question is, which one of the possible approaches is most fair?

The right answer apparently depends on the situation—in particular, how tight the groups are and what contracts have been made between different parties. A service provider with individual customers makes for a relatively loosely coupled group; users within a corporation, on the other hand, are tightly coupled. The importance of Quicksure service may depend on such a non-technical issue as how the service has been marketed, because marketing creates expectations, and expectations have an effect on what is considered fair.

Considerations About Fairness in Reality While not going too deep into the details of service provision, in reality the situation is even more complicated. Several more alternatives emerge if the services form a hierarchy. If Quicksure is high in the service hierarchy, for example, an overload situation of Quicksure service could have different effects on any of the lower-level services. This intricate issue is addressed further in the section titled "Sharing Network Resources Fairly Among All Users" in Chapter 9, "Implementing Differentiated Services."

Another question that might arise is, what is the cohesion among a service group (for instance, among IP telephony users)? A tentative answer is that correlation in behavior may justify the grouping. This is, at least, a reasonable answer from the service provider's point of view because the grouping based on similar behavior may facilitate the network dimensioning and management and by that means improve the cost efficiency of the network. It is not clear, however, whether this is sufficient cohesion to justify any grouping that has considerable effect on the capacity allocation.

The most desirable treatment of a packet depends on the grouping of packets. One more viewpoint can be condensed into the issue of how much an individual packet is responsible for the past traffic process. When a packet arrives at a network node and requires a treatment, for example, the network defines a group of packets for controlling purposes. The

group consists of the last packet with some other earlier packets; that is, the treatment of the packet depends on the arrivals of some previous packets. The earlier packets may belong to a flow of packets generated by an application, to an individual user, or to a larger user group.

The most reasonable and fair approach may considerably differ from case to case. As a consequence, the underlying mechanisms and other building blocks must be able to support different arrangements. That is, network service should be versatile in the sense that it supports purposeful grouping of packets and flows and logical treatment of packets inside the network.

1.2.2 Technical Issues: Versatility and Robustness

Now it is relevant to discuss the most important aspects of technical issues. The next issue to be assessed is the relationship between applications and the network—that is, *network service*. As expressed earlier, the thorough changes in the Internet might be summed up in one word: *fragmentation*—not on the technical level of packet handling, but on the level of applications, users, and business models. If and when the service provider attempts to fulfill all the differing needs, network services must be as *versatile* as possible.

Versatility is, to some extent, an important service attribute for end users as well. Still, it is not clear whether end users want to work with very versatile services and applications, because versatility may bring about complexity and opacity. Hence, versatility is an essential characteristic of the network service to the degree that is necessary for the service provider to offer reasonable service packages for different customers, but it should not be exaggerated by adding insignificant features to customer service packages.

The Internet has relied so far on the benevolence of most end users; that is, the Internet community has been quite a coherent group. Unfortunately, the increased fragmentation of the Internet community brings about an increased threat of undesirable behavior of some end users because of inexperience, greediness, or malevolence. In consequence, it is extremely important to design *robust* network services. Without robustness, it is impossible for the service provider to offer fair and credible customer services. End users must be able to trust that some malevolent or greedy users cannot significantly deteriorate the service of other users.

Three attributes of Differentiated Services have been discussed thus far: fairness, versatility, and robustness. There are, certainly, several other desirable properties, such as reliability, consistency, and simplicity. Although all these could be of primary importance in some cases, they are not used as basic attributes in this book (mainly because they serve some other, more fundamental targets). If the target is making a profitable business, for instance,

simplicity usually means less-expensive management and high reliability means more desirable customer service. This business aspect can be taken into account by an additional attribute, namely, *cost efficiency*.

1.2.3 Cost Efficiency

Efficiency is a common term in the field of communication technology. It usually refers to the relationship between beneficial outcome and the resources spent to realize the desired outcome, such as the average load-to-link capacity ratio. After defining the efficiency by a mathematical formula, it is often thought that the optimal solution is one that just maximizes the efficiency. Although this approach may work well in some cases, it tends to be too limited for practical purposes because it is not reasonable to separate one technical issue from a meaningful framework. As a simple example, just maximizing the number of delivered packets is not a rational target if the cost of the maximization is totally ignored.

Therefore, to emphasize this aspect, the adjunct *cost* is added to the term of efficiency. Cost efficiency refers to the balance of effectively meeting other targets (fairness, versatility, and robustness) at the lowest price; it does not refer to some purely technical issue, such as the number of transmitted packets.

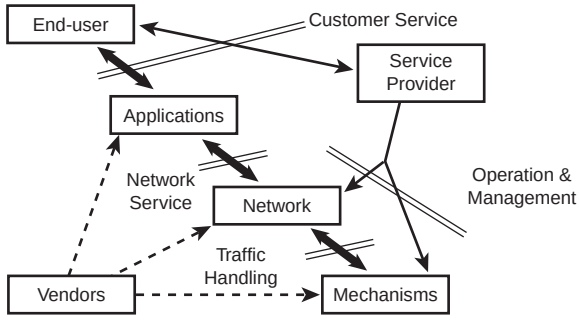
Now it is possible to draw a picture depicting the main entities of Differentiated Services, the main relationships among the entities, and the main targets of the whole project. Figure 1.1 shows the relationships inside the Internet system: customer service between the service provider and the customer, the network service between the network and the application, operation and management between the service provider and the network, and traffic handling between the mechanism and the network.

The main purpose of Figure 1.1 is to illustrate the general structure of Differentiated Services rather than to make any exact statements about the implementation:

- The bold arrows depict the concrete information flow going through the network: Applications send packets into the network, and networks use mechanisms to handle packets in an appropriate manner.
- The standard arrows illustrate the other relationships between entities; for instance, end users use an application through a user interface, and service providers manage the network by using proper tools.
- Double lines illustrate a high-level set of functions, such as services and management. Many of them are somewhat equivocal: It is not clear what issues should be included in customer service or network management. The position of this book, and Figure 1.1 in particular, is that customer service is a broad concept that covers all issues important for customer satisfaction, including the usefulness of applications.

- Broken arrows are used for those interrelations important for real implementations, but of lesser interest in this book.

Figure 1.1 Main entities of the Internet and the relationships among them.



Summary

The technical core of Differentiated Services is the mechanisms used to treat packets in different ways inside network nodes. It is not practical or sufficient, however, to limit this study to the technical level because it does not make it possible to obtain a sufficiently broad view of the primary issues. Therefore, this chapter introduced six other basic entities:

- Service provider
- End user
- Mechanism
- Application
- Network
- Vendor

This basic entity list was introduced mainly to organize the presentation, and it needs further refinement before it is possible to evaluate several complicated issues related to Differentiated Services.

The relationships among these basic entities and the target(s) of their efforts result in Differentiated Services. A target is fixed by four attributes (used extensively in the following chapters):

- Fairness
- Versatility

- Robustness
- Cost efficiency

Throughout this book, the fulfillment of these attributes when using various approaches that have different mechanisms, network services, and customer services is evaluated.

If you look at this introductory chapter, you might notice that it does not have any tight relationship to the Internet Protocol (IP) or the Internet itself. All the basic concepts and ideas can be applied to almost any networking technology based on packets or similar independent information units. It is possible on the one hand, therefore, to apply most of the primary ideas of Differentiated Services to many other networks, such as ATM and Frame Relay; on the other hand, it is possible to exploit the experiences obtained in other networks.