# Challenges in Quality of Service for Tomorrow's Networks

**Luiz A. DaSilva**
*Virginia Polytechnic Institute and State University, USA*

## INTRODUCTION

The original communication networks were designed to carry traffic with homogeneous performance requirements. The telephone network carried real-time voice, with stringent latency bounds, and therefore used circuit-switched technologies with fixed bandwidth allocated to each call. Original data networks were used for electronic mail and file exchange, and therefore employed packet switching and provided best-effort service.

Soon, however, it became clear that economies would accrue from utilizing a common network infrastructure to carry diverse kinds of traffic. Packet switching is now often employed for both real- and non-real-time traffic. This creates a problem: If the same network is to carry flows with diverse performance requirements, it is crucial that it support mechanisms to differentiate among these flows. For instance, real-time applications such as voice and video should be protected when competing for resources with nonreal-time applications such as file transfer or e-mail. Table 1 illustrates some of the quality of service (QoS) requirements of different classes of applications.

The desire to meet the needs of a broad range of applications that coexist in the same network is the primary motivation for the development of QoS mechanisms and architectures.

## BACKGROUND

The concept of QoS arises from the need to allocate resources in order to maximize perceived quality on the basis of intrinsic application requirements (such as low latency for real-time voice), pricing (higher service quality for those willing to pay more), or policy (such as preferred access to network resources for users with mission-critical applications). A discussion of how pricing and policies relate to QoS can be found in DaSilva (2000a) and Flegkas, Trimintzios, and Pavlou (2002), respectively.

Either deterministic or probabilistic QoS guarantees may be associated with a given flow. When the network guarantees that the flow will be allocated bandwidth of at least $B$ bits per second at every router on the source-destination path, it is providing a deterministic guarantee. On the other hand, if at least $p$% of packets belonging to the flow are guaranteed to encounter delay of less than $D$ seconds, the network is providing a probabilistic guarantee.

It is important to note that QoS is not synonymous with performance. In the context of computer networks, the term QoS often implies that some sort of differentiation is made among disparate users. A variety of mechanisms are employed to this end. These include admission control, policing, congestion control, bandwidth reservation, marking, and classification. We briefly discuss each of these mechanisms next.

*Table 1. QoS requirements of different classes of applications*

| Application | QoS Requirements |
|---|---|
| IP telephony | Low delay (on the order of ~ 100 ms) |
| Web surfing | Acceptable throughput |
| Streaming media | Low delay variation (jitter) |
| Networked virtual environments | Low delay in support of interactivity, high bandwidth in support of high-quality graphics |
| Online backup | High throughput, low packet losses (i.e., few retransmissions) |
| E-mail | High tolerance to delay, low to moderate data rate requirements |

The decision of whether to accept a new call or flow is referred to as *admission control* (Breslau, Knightly, Shenker, Stoica, & Zhang, 2000); the objective is to ensure that the network can accommodate all of its current traffic flows with the desired QoS even after the new flow is accepted. In circuit-switched networks, incoming calls that cannot be accommodated are blocked; in packet switching, the flow may be denied access or the packets associated with it may be marked as lower priority and dropped when congestion occurs. This leads to *congestion control*; traffic supported by today's integrated networks tends to be bursty, and the admission decision is generally not made based on peak traffic conditions. It is therefore possible, even after admission control, that the network may experience congestion at times. Frost (2003) presents a study of the effects of temporal characteristics of congestion on user-perceived QoS. Measures to alleviate congestion include the dropping of packets at congested routers, as well as implicit or explicit signaling to the source to reduce its transmission rate. These measures may take into account QoS requirements, for instance, by requesting that one source reduce its transmissions while another (more critical) source is allowed to maintain its current rate.

*Policing* refers to measures taken by the network to ensure that the traffic being offered by a user conforms to a preagreed traffic contract. Excess traffic can be simply dropped at the ingress router or marked for best-effort delivery. Conversely, users may employ *shaping* to ensure their traffic conforms to preestablished parameters such as maximum data rate or maximum burst length. *Bandwidth reservation* may be used in packet-switched networks to provide minimum guarantees as to bandwidth availability to a flow. This requires a signaling phase to precede the transmission of packets, during which each router on the source-destination path agrees to reserve a portion of the available bandwidth to be used by the flow. Queuing and scheduling mechanisms such as weighted fair queuing are implemented by routers in order to meet such guarantees. Unlike in circuit switching, reserved bandwidth that is not being used by the reserving flow is generally made available to other flows through the use of work-conserving scheduling. To support service differentiation, packets are often *marked* using preassigned bit sequences in the packet header; this allows routers in the path to recognize the packet as part of a given flow and *classify* it accordingly (Gupta & McKeown, 2001).

QoS mechanisms can be provided at different layers of the protocol stack as well as by the application and the middleware (DaSilva, 2000b). At the physical and data link layers, prioritization, forward error correction, code, and slot assignment can be adopted for service differentiation. For instance, in random-access local area networks, nodes must back off in case of collision, picking an interval before they are allowed to attempt retransmission; by enforcing different back-off intervals for different nodes, we can achieve prioritization in access to the channel at times of heavy traffic. Scheduling, shaping, admission, and flow control are some of the mechanisms described above that may be adopted at the network layer. Middleware is sometimes developed to take care of classification of flows and marking of packets, and generation of resource-reservation requests. The application itself may employ prefetching and caching of information to improve performance experienced by selected users.

Asynchronous transfer mode (ATM) is one mature example of a packet-switched network providing QoS differentiation. In ATM, this is achieved by defining multiple service categories with associated QoS guarantees and traffic conformance definitions (Giroux & Ganti, 1999). Due to the ubiquity of the Internet protocol (IP) and current interest in real-time traffic using this protocol (voice over IP, for instance), recent research has focused on how to support QoS over the Internet (Armitage, 2000; Firoiu, Le Boudec, Towsley, & Zhang, 2002).

## FUTURE TRENDS

An important challenge in providing QoS differentiation in packet-switched networks has to do with the *scalability* of such mechanisms (Welzl & Muhlhauser, 2003). Stringent QoS guarantees require that all network nodes store information about individual flows in order to make scheduling decisions. While this is reasonable in an intranet, with a limited number of flows and complete control over the network by a single administrative unit, those types of approaches do not scale well. In particular, in the Internet, a core router may be routing packets belonging to many thousands of flows at any one time, and maintaining state information about each flow is infeasible. Stateless approaches achieve better scalability: They classify packets into a finite, reasonably small set of classes, marking each packet accordingly, and associate probabilistic service guarantees with each class. The Internet Engineering Task Force (IETF) has been studying QoS architectures for the Internet for several years. Any major change to the IP suite is, of course, always controversial (one must only look at the time it is taking for the widespread adoption of IPv6 for another example of the resulting inertia). While several important developments resulted from these working groups (Blake, Black, Carlson, Davies, Wang, & Weiss, 1998; Zhang, Deering, Estrin, Shenker, & Zappala, 1993), the ultimate goal of having QoS widely available in the Internet remains elusive.

Providing QoS guarantees in mobile wireless environments is another great challenge. The wireless medium is