OVERVIEW OF QOS

Two broad approaches to QoS are integrated services and differentiated services. Integrated services provide QoS to individual applications and flow records. Providing QoS requires certain features to be maintained in switching nodes. QoS protocols govern traffic shaping and packet scheduling. Traffic shaping regulates the spacing between incoming packets. Other advanced QoS protocols covered are admission control and RSVP.

The differentiated services provide QoS support to a broad class of applications. The basics of resource allocation for packet-switched networks are reviewed, as is resource allocation for all possible layers of the protocol stack. Resource-allocation algorithms can be used to avoid possible ATM congestion.

Communication networks face a variety of quality-of-service demands. The main motivation for a QoS unit in a data network port processor is to control access to available bandwidth and to regulate traffic. Traffic regulation traffic is always necessary in WANs in order to avoid congestion. A network must be designed to support both real-time and non-real-time applications. Voice and video transmissions over IP must be able to request a higher degree of assurance from the network. A network that can provide these various levels of services requires a more complex structure.

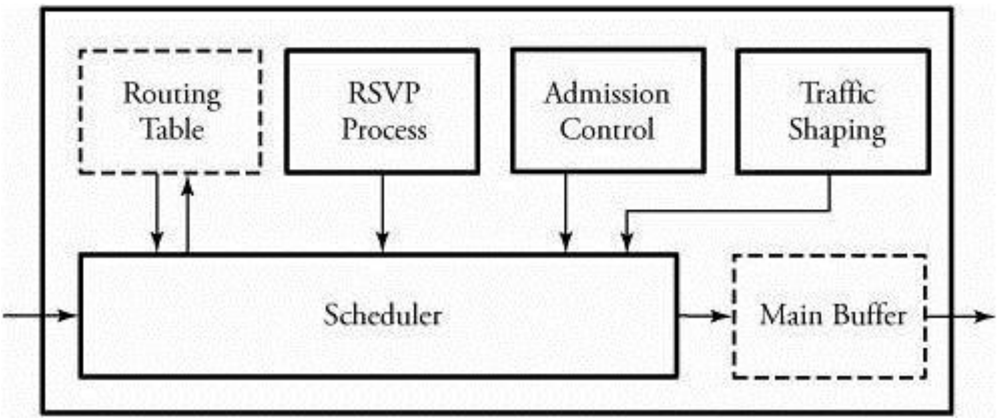
The provision of QoS to a network either does or does not come with a guarantee. Nonguaranteed QoS is typically based on the best-effort model, whereby a network provides no guarantees on the delivery of packets but makes its best effort to do so. In a non-real-time application, a network can use the retransmit strategy for successful data delivery. However, in a real-time application, such as voice or video networking, where timely delivery of packets is required, the application requires a low-latency communication. Consequently, the network must be able to handle packets of such applications more carefully.

Integrated Services QoS:

The integrated services approach, consisting of two service classes, defines both the service class and mechanisms that need to be used in routers to provide the services associated with the class. The first service class, the guaranteed service class, is defined for applications that cannot tolerate a delay beyond a particular value. This type of service class can be used for real-time applications, such as voice or video communications. The second, controlled-load service class, is used for applications that can tolerate some delay and loss. Controlled-load service is designed such that applications run very well when the network is not heavily loaded or congested. These two service classes cover the wide range of applications on the Internet. A network needs to obtain as much information as possible about the flow of traffic to provide optimum services to the flow. This is true especially when real-time services to an application are requested. Any application request to a network has to first specify the type of service required, such as controlled load or guaranteed service. An application that requires guaranteed service also needs to specify the maximum delay that it can tolerate. Once the delay factor is known to a service provider, the QoS unit of the node must determine the necessary processes to be applied on incoming flows. Figure 6.2. shows four common categories of processes providing quality of service.

1. Traffic shaping regulates turbulent traffic.
2. Admission control governs whether the network, given information about an application's flow, can admit or reject the flow.
3. Resource allocation lets network users reserve bandwidth on neighboring routers.
4. Packet scheduling sets the timetable for the transmission of packet flows. Any involving router needs to queue and transmit packets for each flow appropriately.

Figure 6.2. Overview of QoS methods in integrated services



The widespread deployment of the integrated services approach has been deterred owning to scalability issues. As the network size increases, routers need to handle larger routing tables and switch larger numbers of bits per second. In such situations, routers need to refresh information periodically. Routers also have to be able to make admission-control decisions and queue each incoming flow. This action leads to scalability concerns, especially when networks scale larger.