Internet services beyond best effort

The packet-based data transmission of the Internet allows the multiplexing of a variety of simultaneous connections originating from sources with different characteristics (e.g. voice, video, data). Traditional Internet is based on a best effort service, whereas future Internet services make more and more special demands on the communication network. This paper presents a general overview of fundamental mechanisms for guaranteeing Quality of Service (QoS) in the current and future Internet. After introducing general QoS concepts, basic mechanisms in network routers are investigated. Furthermore, the two basic architectures of Integrated Services (IntServ) and Differentiated Services (DiffServ) are discussed in detail. Finally, the analytical investigation of a special Voice-over-IP scenario demonstrates the applicability of the relevant QoS concepts and their positive consequences with respect to the quality of this voice service. A short overview of current hot research topics concludes the paper.

Keywords: quality of service; Internet; TCP/IP; integrated services; differentiated services; voice over IP; resource reservation; admission control

1. Introduction

The basic need for Quality of Service (QoS) comes from the fact that any arrangement of providing a fixed one-size-fits-all service to all customers is insufficient. Users employ different applications, for example Voice over IP (VoIP), Virtual Private Networks (VPNs), Web-browsing, and mission-critical e-commerce, causing the need for a variety of services to be provided by the network. Different services require different treatment of data packets in the network. Hence, QoS is a term referring to technologies that classify network traffic and ensure that some of that traffic receives preferential handling. Customers will be enabled to subscribe different services in terms of throughput, loss, response time, delay and/or delay jitter (the so-called QoS parameters) and will be charged for the actual level of service provided. Literature about QoS can be found in (Ferguson, Huston, 1998; Siegel, 2000; Greenville Armitage, 2000; Uless Black, 2000; Kilkki, 1999).

The implementation of the QoS concept began with the emergence of X.25 and its improvements in the late 1960s and early 1970s. The goal was to create a common platform on which all users with a special interface could participate with their data traffic. The design of X.25 allowed users to request certain levels of service, for the first time.

The next steps in providing QoS were the introduction of the Frame Relay (FR) and Asynchronous Transfer Mode (ATM) technologies. FR gained much popularity as a method for interconnecting widely separated geographic locations, but has limited QoS capabilities. ATM was planned to transport as different traffic types as voice, data, video linked with QoS guarantees in a single generic technology. Because ATM was expected to solve many problems in delivering different services, it has seen extensive specifications, standards and research. ATM, however, suffers from the fact that specifications were only partially implemented. Originally, the Internet Protocol (IP) offered only one service - “best effort”. Best effort means that each user is provided service and gets a fair share of the available network resources. On the other hand, no promises can be made concerning QoS guarantees like an upper bound on end-to-end delay, minimum throughput and delay jitter.

Internet traffic is solely controlled by congestion control mechanisms at hosts. Congestion control adjusts the sender’s transmission rate into the net as a function of the network load.

PLASSER, Erich, Dipl.-Ing., DWT, ZIEGLER, Thomas, Dipl.-Ing. Dr., REICHL, Peter, Dipl.-Math. Dr., MITTENECKER, Georg, Dipl.-Ing., ifw. Forschungszentrum Telekommunikation Wien, Techgate, Donau-City-Straße 1/9, A-1220 Wien.
if the network load is high, and the available capacity diminishes, the sending rate is reduced. If the network load is low, the sending rate is increased correspondingly. Congestion control is both a means to avoid congestion collapse and to guarantee high network utilization. For the Internet, the Transmission Control Protocol (TCP) congestion control (Jacobson, 1988; Stevens, 1994) is of central importance. TCP congestion control uses packet loss as an indication to congestion and adjusts a window of packets as a function of the number of lost packets during the last round trip time. The effective TCP sending rate can be estimated as window size divided by round trip time.

Such an architecture has significant technical advantages in terms of simplicity and high link utilization, is perfectly suited for classical Internet applications like e-mail, FTP (File Transfer Protocol) and Telnet, but lacks the facilities to provide service guarantees. Although a rudimentary QoS support (Type of Service bits in the IP header) was intended from the very beginning, this was hardly used in actual implementations. Nowadays users are no longer satisfied with simple applications like FTP or Telnet alone. Some of the new Internet applications require significant bandwidth. Others, like IP telephony and other real-time applications have strict constraints on timing requirements. These applications require network services beyond the simple "best effort" service that IP delivers.

2. Basic QoS mechanisms in IP

The goal in a QoS-enabled environment is to allow for predictable service delivery to certain classes or types of traffic regardless of other traffic flowing through the network at any given time. We describe the accomplishment of a QoS-connection to achieve this goal in a very generic way by emphasizing on two fundamental steps:

1) Admission control and resource allocation:

Before the user may send data packets the network has to perform admission control. Given a certain load situation and QoS demand of a flow, admission control determines whether the flow's demand can be granted without violating the service guarantees of existing flows. The decision about admission can be made anywhere in the network. Nevertheless, the enforcement about admission control is delegated to a policer at an edge router (Fig. 1). If the flow is admitted it shares the available network resources with many other flows. Thus it is important to allocate to each flow or each group of flows resources according to predefined QoS guidelines. In most cases resource allocation is made for bandwidth and buffer space, enabling the network service provider to deliver QoS guarantees.

2) Maintaining the arranged network resources during the lifetime of a flow:

During the packet flow it is to assure that the adjusted conditions are kept and the flow's parameters are maintained by the router components: classifying, policing, queue management, scheduling.

The following section explains this second step in more detail.

3. Router components in QoS-enabled networks

The routers are the elements in a network where resources (bandwidth, buffer space) are shared by the different traffic types and where algorithms can be adjusted according to the expected different handling. A traditional router mainly focuses on where to send packets. It makes forwarding decisions based on the destination address in each packet and locally held forwarding information tables. QoS-enabled routers must additionally provide control of when to send packets, which packets to send and which packets to drop in case of congestion. Figure 2 shows the components in a router necessary for QoS handling which are part of a router's forwarding process. The following subsections address these and several other associated router components.

Fig. 2. Forwarding process within a router

3.1 Classification

Receiving a packet, a classifier determines which traffic class the packet belongs to based on the content of one or several fields in the packet header. If several fields are needed, this process is called Multifield Classification (MF). Classifiers are used to pass packets matching some specified rule to a traffic conditioner (which is a term for metering, shaping, policing) (Blake et al., 1998) as well as to pass packets to a specified queue management mechanism. Figure 3 shows how a classifier assigns different packets to different queues.

Fig. 3. Router output port architecture

3.2 Policing and marking

In most cases policing is about the supervision of whether the incoming traffic is compliant with pre-negotiated policies (then it is called traffic policing). In general the term policing can be applied to a broad range of rule systems. The term also appears at shaping, dropping, authentication, billing and may involve two (bilateral) or more (multilateral) parties.

Every traffic class has certain limits to its allowable temporal behavior – for instance, a rate limit specified by the number of packets per second.