Supporting Quality of Service for Internet Multimedia Applications

Yew-Hock Ang
Nanyang Technological University, Singapore

Zhonghua Yang
Nanyang Technological University, Singapore

INTRODUCTION

The Internet has gone from near-invisibility to near-ubiquity and penetrated into every aspect of society in the past decades (Department of Commerce, 1998). The application scenarios have also changed dramatically, and now demand a more sophisticated service model from the network. In the early 1990s, there was a large-scale experiment in sending digitized voice and video across the Internet through a packet-switched infrastructure (Braden, Clark, & Shenker, 1994). These highly-visible experiments have depended upon three enabling technologies: (1) Many modern workstations now come equipped with built-in multimedia hardware, (2) IP multicasting, which was not yet generally available in commercial routers, and (3) Highly-sophisticated digital audio and video applications have been developed. It became clear from these experiments that an important technical element of the Internet is still missing: multimedia, which dominate increasing proportion of today's data traffic, are not well supported on the Internet.

BACKGROUND

The Internet, as originally conceived, offers only best-effort service provided by the connectionless datagram service of IP (Internet Protocol). Such service does not promise whatsoever the end-to-end delay of packets nor about the variation of packet delay variation, and is not suitable for carrying multimedia traffic. To support multimedia services, the Internet must guarantee its transmission delay, delay variation, and loss according to the quality of service (QoS) required by multimedia applications and the service quality level expected by the Internet users. In other words, the Internet is to be QoS-aware of different service quality requirements for supporting multimedia applications and sensitive to the characteristics that are peculiar to multimedia data.

On the other hand, traditional networks are QoS-specific networks which were designed to support fixed classes of service; such as telephone networks for time-sensitive voice communication, cable networks for error-tolerant video streaming, and data networks for non real-time error-sensitive text messaging. The future Internet, however, is expected to be QoS-aware, and provide QoS guarantees to meet a wide range of QoS requirements for supporting multicast multimedia applications. In response to these expectations, the Internet Engineering Task Force (IETF) has specified three service models for supporting QoS on the Internet; namely, the Integrated Services (IntServ), Differentiated Services (DiffServ), and Multi-Protocol Label Switching (MPLS). This article discusses the design motivation, architecture and development of these service models for supporting QoS guarantees on the Internet.

PRINCIPLES FOR SUPPORTING MULTIMEDIA QOS

To support multimedia applications, the Internet service models must address the issue of quality of service that relates to the characteristics of multimedia data. These data characteristics are: (a) media type, (b) data synchrony, and (c) data persistency. The media type identifies the multimedia as one of image, video, audio, or text. Network services should be aware of different error susceptibilities of media types. For example, video data are handled by a QoS-aware network with a higher packet loss rate or treated with higher drop precedence than for error-sensitive audio data. Data synchrony defines the temporal synchronization of real-time data, such as interframe delay in video streams. A QoS-aware network must forward real-time data packets with higher priority than for non real-time data packets. Data persistency describes the transient nature of "live" data. Live data are not stored or buffered at its source; hence they are nonpersistent and are not retransmitted when received in error.

To be sensitive to QoS requirements of multimedia applications, a QoS-aware network must be designed according to the following four service principles and implemented with a corresponding set of traffic management/handling mechanisms:

1. *Service agreement*: a call admission mechanism to allow an application to negotiate for a service agreement

to share network resources or to operate at a required service level.
2. *Service specification*: a packet marking mechanism for the user or application to specify the application data type and a classification mechanism to allow routers to distinguish between different data specifications or classes of applications.
3. *Service conformance*: a traffic conditioning mechanism for flow policing and remarking of misbehaved packets so as to force application traffic to conform to service agreements.
4. *Service commitment*: a queue management and scheduling mechanism for QoS treatment of packets, while ensuring high resource utilization, to meet service agreements, and thereby fulfilling service commitments.

These service principles specify the manner a network is made aware of the service level requirements, specifications of traffic it is expected to handle, policing policies for service agreement enforcement, and service commitment to guarantee QoS of admitted flows. Accordingly, future Internet service models must all implement these principles for QoS-awareness.

THE INTERNET INTEGRATED SERVICES (INTSERV) MODEL

The Integrated Services (IntServ) model extends the original connectionless point-to-point best-effort service of the Internet with a connection-oriented end-to-end QoS provisioned service (Braden et al., 1994). It achieves this by setting up a routed path between the endpoints with the necessary level of resources reserved along this path to guarantee end-to-end QoS on a per-flow basis (Figure 1).

Implementation of QoS support mechanisms for QoS guarantees in IntServ:

1. *Service agreement*: A call setup, using RSVP signaling (Braden, Zhang, Berson, Herzog, & Jamin, 1997) initiated by the application, requests for explicit resource reservation in routers to satisfy per-hop QoS requirements of an anticipated flow. The QoS requirements specified in a PATH message carries the *Sender_TSpec* and the *AdSpec* from the source toward the receiver. The *TSpec* contains the description of the traffic profile (traffic burst rate and average rate), while the *AdSpec* describes the properties of the data path and QoS requirements (packet delay and loss) of the sending application. The receiver responds to the service request with a RESV message for setting up the resources in the routers along the data path toward the sender.
2. *Service specification*: The RESV carries a *Flowspec* and a *Filterspec*. The *Flowspec* contains a *Receiver_TSpec* which describes the traffic profile, and *RSpec* (Reserved QoS spec) the QoS service level that each router along the routed path is expected to provide. The *Filterspec* provides the information required by the packet classifier in the router to identify packets belonging to a particular flow. Such flow-state information must be maintained in end-hosts and routers along the routed path for flow filtering and QoS packet treatment.
3. *Service conformance*: Packet classifiers in routers filter incoming flow packets based on the *Filterspec*. Packets identified as belonging to a provisioned flow are treated according to the reserved level of QoS. Flow packets not identified are treated on a best-effort basis.
4. *Service commitment*: Packet schedulers in routers handle a filtered flow according on its *TSpec* and service it such that its QoS commitments (*Rspec*) are met. The Token Bucket scheduler is typically used to service a flow carrying a *TSpec* of traffic burst rate, B bounded by the token bucket size, b; and its long-term traffic average rate, R bounded by the token arrival rate, r. The flow delay defined by b/R is guaranteed so long as the burst rate, B does not exceed the bucket size, b.

Figure 1. End-to-end QoS provisioning in Integrated Services Model