STARDUST® forums™

Stardust Forums, Inc.
1901 S. Bascom Ave, Suite 333
Campbell, California 95008
Phone: 408-879-8080
Fax: 408-879-8081
www.stardust.com
www.ipmulticast.com

White Paper -
Internet Bandwidth Management

Raw data speed is important, but it isn't enough
The Internet's data throughways need control

Internet Bandwidth Management

Raw data speed is important, but it isn't enough

The Internet's data throughways need control

Internet Bandwidth Management

Raw data speed is important, but it isn't enough
The Internet's data throughways need control

Scope of this document

In the early days of the Internet Engineering Task Force [IETF], a popular T-shirt said "IP over everything." This catchy phrase characterized the ability of the Internet Protocol [IP] to run across virtually any network transmission media and communicate between virtually any system platforms. It was to a large extent this flexibility that led to IP's phenomenal success.

With the tremendous growth of the Internet in the past few years, and the wide variety of new applications that have appeared, the convergence of other networks--telephone, radio, and television--to the Internet is underway. "Everything Over IP" could be the T-shirt's new incarnation. But this trend is not without strain.

Network traffic has increased as the number of users and applications has increased, obviously. And Moore's Law [Moore] keeps producing faster computer systems capable of transferring more data than ever. The question is whether increasing bandwidth--the data carrying capacity of the network--is sufficient to accommodate these increased demands. The answer is no, it is not. Internet traffic has not only increased, but it has changed in character. New applications have new service requirements, and as a result the Internet needs to change as well. In this paper we explain why and how the Internet must change to satisfy the needs of these new applications.

We begin with a quick review of the design characteristics of IP that led to its success. We then describe the new breed of Internet applications, and examine their network requirements. We survey the many new bandwidth technologies now available, then explain why they are insufficient to provide the whole solution. In the remainder of the paper we describe why bandwidth management is the answer, describe the technologies that enable it, their architecture and support services they require. We describe some of the business opportunities these bandwidth management services create, and wrap-up with a vision of future Internet applications that will result.

Introduction

The Internet Protocol (IP) has enabled a global network between an endless variety of systems and transmission media. Around the world email exchange and web browsing are a part of everyday life for work, study and play. And by all indications other networks--phone, radio, and television--are also converging on IP to leverage its ubiquity and flexibility. With these new networks come new applications ...and more new users. There's no sign that the phenomenal growth of the Internet will subside any time soon.

One reason for IP's tremendous success is its simplicity. The fundamental design principle for IP was derived from the "end-to-end argument" [e2e], which puts "smarts" in the ends of the network--the source and destination network hosts--leaving the network "core" dumb. IP routers at intersections throughout the network need do little more than check the destination IP address against a forwarding table to determine the "next hop" for an IP datagram. If the queue for the next hop is long, the datagram may be delayed. If the queue is full or unavailable, an IP router is allowed to drop a datagram. The result is that IP provides a "best effort" service that is subject to unpredictable delays and data loss.

Bandwidth is the Answer! What's the Question?

As a result of the tremendous growth of the Internet, IP's weaknesses are showing. Increasing the available bandwidth to avoid congested Internet links is the obvious solution. But the problem is more than a simple capacity issue.

The issue is that not only has traffic increased in volume, it has also changed in nature. There are many new types of traffic, from many new IP-based applications, and they vary tremendously in their operational requirements.

The Changing Needs of Internet Applications

Some of the new breed of Internet applications are multimedia and require significant bandwidth. Others have strict timing requirements, or function one-to-many or many-to-many (multicast). These require network services beyond the simple "best-effort" service that IP delivers. In effect, they require that the (now "dumb") IP networks gain some "intelligence"

IP Telephony is today's "killer application." More than any other, the desire to provide telephone service over the Internet is driving the convergence of the telephone and Internet industries. This is quite interesting, since the design principles behind the telephone networks are almost exactly the opposite as those behind IP networks. Whereas IP uses (datagram) packet-switching and provides best effort services, telephone networks use (connection-oriented) circuit-switching to provide provisioned service. There's a reason for this: A two-way, real-time telephone conversation is a demanding application for a network to satisfy.

Applications: Raising the Bar

Networks exist to support applications. As a result, applications spur network advances. Or more precisely, application *users* drive them.

Different network applications have different operational requirements that demand different network services. Increased network traffic requires increased network bandwidth capacity, but new applications like IP telephony have other requirements, and increasing the network "pipe-size" is not the only answer.

Rate Type	*Descriptions*
Stream	Predictable delivery at a relatively constant bit rate (CBR). For example, although their rates often fluctuate, audio and video data streams are considered CBR because they have a quantifiable upper bound.
Burst	Unpredictable delivery of "blocks" of data at a variable bit rate (VBR). Applications like file transfer move data in bulk that can increase data rate to use all available bandwidth (no upper bound).

Table 1: Terms to characterize application data rates in terms of relative predictability.

Delay Tolerance	*Delivery Type*	*Description*
high	**Asynchronous**	No constraints on delivery time (a.k.a. "elastic")
	Synchronous	Data is time-sensitive, but flexible.
	Interactive	Delays may be noticeable to users/applications, but do not adversely affect usability or functionality.
	Isochronous	Time-sensitive to an extent that adversely affects usability.
low	**Mission-Critical**	Data delivery delays disable functionality.

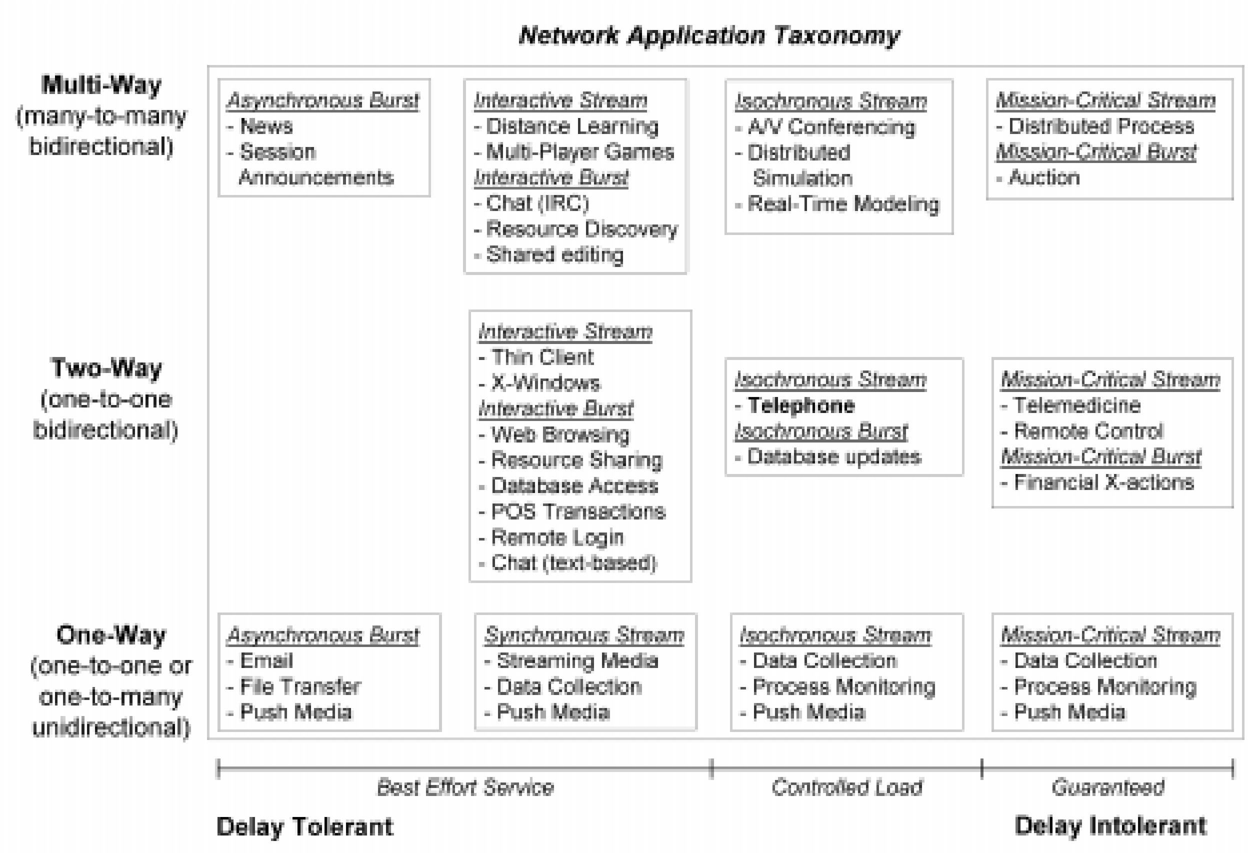
Table 2: Terms to characterize application sensitivity to data delivery delays.

Network applications can be characterized in terms of how predictable the data rate is (see Table 1), and how tolerant of delay delivery is (see Table 2). Generally, two-way applications are more sensitive to delay than one-way, as we describe next and illustrate in Figure 1.

The Demands of Convergence: Multimedia and Multicast

For data networks, the convergence to IP is basically a done deal. Many information technologies and commerce applications have already moved to IP, and others are en route (though some will always remain on legacy networks). Most applications are low-priority and low-bandwidth, with a high tolerance for delay, but others have strict operational requirements.

For radio and television networks, the convergence on IP has started but still has a ways to go. First and foremost, they need bandwidth, and that is coming (as we describe shortly). The broadcast models in existence today are a close match to the one-to-many IP multicast model. Broadcasters currently service millions of customers simultaneously, and unicast on the Internet could never possibly scale to these levels. Hence, multicast deployment is necessary to enable the convergence of television and radio networks to IP.



Multi-Way (many-to-many bidirectional)	*Asynchronous Burst* - News - Session Announcements	*Interactive Stream* - Distance Learning - Multi-Player Games *Interactive Burst* - Chat (IRC) - Resource Discovery - Shared editing	*Isochronous Stream* - A/V Conferencing - Distributed Simulation - Real-Time Modeling	*Mission-Critical Stream* - Distributed Process *Mission-Critical Burst* - Auction
Two-Way (one-to-one bidirectional)		*Interactive Stream* - Thin Client - X-Windows *Interactive Burst* - Web Browsing - Resource Sharing - Database Access - POS Transactions - Remote Login - Chat (text-based)	*Isochronous Stream* - **Telephone** *Isochronous Burst* - Database updates	*Mission-Critical Stream* - Telemedicine - Remote Control *Mission-Critical Burst* - Financial X-actions
One-Way (one-to-one or one-to-many unidirectional)	*Asynchronous Burst* - Email - File Transfer - Push Media	*Synchronous Stream* - Streaming Media - Data Collection - Push Media	*Isochronous Stream* - Data Collection - Process Monitoring - Push Media	*Mission-Critical Stream* - Data Collection - Process Monitoring - Push Media

Figure 1: Application Taxonomy in terms of data distribution and time constraints. Notice that the placement of some applications--like Push Media--are content dependent (for example, stock quotes are time-sensitive, whereas headline news and weather are not)

The added value that IP networks can provide for audio/video applications is tremendous. They enable new dimensions to multimedia content. They can embed web links, or simultaneously send slides and files or other content during transmission, to enrich the delivery. They enable two-way communications, so content receivers can "talk-back" to providers. And since multicast allows receivers to talk to other receivers as well (i.e. many-to-many), they open the door to all-new application possibilities (discussion groups, interactive distance-learning, game-shows, instant surveys, etc.). They provide the potential for global audiences, which can make it economically feasible to provide niche focus.

But before this tremendous potential can be fulfilled, there are other network service requirements to satisfy in addition to multicast.

Real-Time and Two-Way

As mentioned earlier, IP telephony is today's premier killer application. The market pressure to enable IP telephony has done more to expose the service deficiencies of IP, and upped the ante to define standards and deploy managed bandwidth on IP networks. Although it is a multimedia (audio) application, its bandwidth requirements are relatively modest (about 8Kbps each way), so bandwidth is not the issue. Latency is.

For IP telephony--and other real-time or two-way applications--the timing requirements are much more significant than the bandwidth requirements. There's a person at either end of the conversation, and they have immediate and obvious evidence of the quality of a call--or lack of it. Dropouts and delays are noticeable and distracting. Round-trip delivery delays above 0.5 second can make them unusable.

The current consumer standard for usability is the cell-phone. As anyone who has used one knows, they are not perfect. Noise and dropped-calls are not uncommon. On the other hand, latency has never been an issue. Thus, despite it's short-comings, cell-phone service is considered far better than what is typically possible using "best effort" IP service over the standard Internet. This illustrates the impact on usability that traffic delays represent.

Routing delays and lost packets due to transient network congestion--from the unpredictable, bursty nature of network traffic--results in sub-optimal round-trip times on IP networks that severely limit the usability of IP-based telephone service. "The telecommunications industry started with a specific application (i.e. telephony) and built a network to suit it. The Internet, on the other hand, started in exactly the opposite way: it started with a new network technology and explored, successfully, new applications that were able to use the undefined (best-effort) service" [Paradigm].

Increasing bandwidth capacity will improve IP service, so it is an essential first step towards solving the latency issue. However, it is not enough to satisfy telephony application requirements.

A Spectrum of Possibilities

Bandwidth is the raw data carrying capacity of a network, the resource by which we most commonly gauge a network's capabilities. It is a measure of how many elemental bits of information a network can move from one host to another in a unit of time (seconds) under ideal conditions. Unfortunately, most networks are far from ideal. And in terms of consistent bandwidth capacity between any two hosts, the public Internet is ideal less often than most.

The Internet is a network of networks, a mesh of various transmission media, with a wide range of bandwidth capacity and latency characteristics. Link status between two hosts across the Internet can vary widely from one millisecond to the next. Network application traffic is bursty, and with many applications sharing the same network links at the same time, transient congestion is often the result. Congestion causes delivery delays or data-loss.

This has not been a problem for delay-tolerant (i.e. "elastic") applications like email, nor even for interactive file transfer and web-browsing. But delays can be fatal to mission-critical applications, or significantly limit the usability of real-time applications like telephony.

Bandwidth is a Requirement

Internet growth means more hosts, networks, users and applications. It manifests as increased 7x24 network traffic of a wider variety. Hence, the bandwidth needs of the Internet are ever increasing. Adding bigger, faster network pipes is a necessity, not a luxury.

Bandwidth is Inevitable

Fortunately, Moore's Law helps in this regard by (indirectly) fostering new, and ever-faster bandwidth technologies for long, medium and short-haul (WAN, MAN and LAN). As Andrew S. Grove, Intel's chairman and co-founder, said in Forbes ASAP magazine, "if you are amazed by the fast drop in the cost of computing power over the last decade, just wait till you see what is happening to the cost of bandwidth." [Grove]

The debate is not whether bandwidth will be available, but what the technologies will be. There are many in various stages of development and deployment, the most exciting of which are in the realms of broadband wireless and low-orbit satellite. The technologies in use today rely on fiber- optics, geostationary (high-orbit) satellite, coax, utp, cat5 and standard telephone copper. The current breed of high-bandwidth protocols includes ATM (Asynchronous Transfer Mode), Gigabit Ethernet, Frame Relay, SONET (Synchronous Optical Network), WDM (Wavelength Division Multiplexing, SMDS (Switched Multi-megabit Data Service), and xDSL (Digital Subscriber Line). The race is on to deploy and market these technologies to businesses and consumers, and it's moving quickly.

Interoperability is Key

Since the world is moving en masse to the Internet Protocol, interoperability between IP and these high-speed communications technologies and protocols is necessary. IP is media independent, so typically it is not a problem. However, the operational characteristics of some media do not map well to the IP mechanics. For example, satellite is often one way and some protocols--like TCP--rely on two-way. Even when satellite does have a back-channel, it may be asymmetric (e.g. on a dial-up). The fact that satellite provides very high bandwidth coupled with high latency also raises issues that need to be dealt with. But these issues have been identified and addressed to a large extent. A significant advantage to satellite is its perfect IP multicast support, which minimize its multicast deployment challenges.

Of particular importance is ATM (Asynchronous Transfer Mode), the darling of the telephone industry. ATM and IP have to "play nice" together if the telephone network is to interoperate (and converge) with the Internet to enable seamless telephony services.

ATM plays an important role in telephone network backbones, and its salient feature is "quality of service" (QoS) support. By allocating resources to a virtual circuit during connection setup that remain dedicated for the duration of the connection, ATM can satisfy the real-time (isochronous) delivery requirements of a two-way phone conversation.

The virtual circuit architecture of ATM is in stark contrast to the packet-switched design of IP, however. In addition to ATM's 53-byte "cell" size, and the fact that ATM is a data-link layer protocol as well as a network-layer like IP, these differences raise questions about compatibility.

Fortunately, the work to ensure that IP can operate over ATM networks is done, and proven to work well. Even the unlikely mapping of IP multicast to ATM multipoint is shown to operate effectively [Maufer]. ATM's Available Bit Rate (ABR) service is actually intended to provide a service similar to IP's Best Effort.

Current market indications are that ATM is unlikely to go to the desktop (i.e. to Internet hosts). This means that despite ATM's provisioning, at least part of the network will still only provide IP's "best effort" service. Hence, they are still susceptible to network congestion that can affect data throughput, and thereby adversely affect a telephony application. IP networks need a way to map to the QoS of ATM and extend it to the pure-IP portions of the Internet. To this end, work is underway to map ATM QoS to IP's RSVP (which we describe later).

Necessary but Insufficient

Another popular T-shirt among IETF engineers says "IP: necessary and sufficient." The obvious implication is that other network protocols are superfluous, and IP's best effort service can satisfy any application's requirements.

This is true, assuming the network's bandwidth capacity is sufficient to avoid any delays or dropped datagrams. But as anyone who uses the Internet knows, network delays are a common occurrence. Internet traffic increases in proportion to available bandwidth as fast as it is added, so delays are inescapable. And then there are times when traffic increases to extraordinary proportions--such as the release of the Starr Report about the Clinton-Lewinsky affair, when Microsoft released version 3 of its Explorer browser update, and after the Hale-Bopp cult suicides. In all these instances, resulting traffic caused congestion that effectively disabled sections of the Internet. When this happens, IP's best effort--which provides uniform service levels to all users--is uniformly bad.

Over-Provisioning is Unrealistic

The obvious solution to handle these peak periods is to over-provision the network, to provide surplus bandwidth capacity in anticipation of these peak data rates during high-demand periods. Equally obvious, however, is that this is not economically viable--at least not with today's bandwidth technologies and infrastructures. Since peak data rates and the network regions on which they might occur are seldom possible to predict, this is not a realistic alternative anyway.

IP is necessary and bandwidth is necessary, but neither is sufficient for all application needs under all conditions. Best effort cannot always provide a usable service, let alone an acceptable one. Even on a relatively unloaded IP network, delivery delays can vary enough to adversely affect applications that have real-time constraints. To provide service guarantees--some level of quantifiable reliability--IP services must be supplemented. And this requires adding some "smarts" to the net that can differentiate traffic and enable different service levels for different users and applications. In other words, IP networks need active bandwidth management.

Bandwidth Management Technologies

The purpose of a network is to service network applications. The goal of Bandwidth Management is to make a concerted effort within the network elements to differentiate between applications and address their needs accordingly. It also means using the network elements efficiently, distributing traffic to avoid congestion points.

There are many mechanisms used to enable bandwidth management, and they range in complexity, granularity and speed, all of which correlate. These mechanisms may be passive--simply dropping packets--or explicit, such as congestion notification, resource reservation, traffic classification, packet marking, policing (packet drop algorithms), queue distribution, and admission control.

The passive mechanisms have always existed in the IP best-effort networks, and the active mechanisms are new and controversial. They involve monitoring and controlling packet queues within network elements (e.g. in IP routers). They also assume the existence of other new services such as policy management, authentication, and accounting infrastructure, as well as application programming interfaces (APIs) to allow application control and feedback.

Passive Bandwidth Management

From a network standpoint, bandwidth management is implicit in a best-effort IP network. After a network element receives a packet and determines the outgoing interface to forward it to, it may find the outgoing queue for that interface is currently full. This occurs when network traffic exceeds the network element's bandwidth capacity--it is congested--and as a result it drops the packet silently [SrcQuench].

TCP (Transport Control Protocol) is a reliable transport that uses acknowledgements (ACKs), and when a send timeout occurs or duplicate ACK is received--either of which indicate a lost packet--it assumes the network is congested and engages a back-off algorithm to slow down the send rate. It also has a slow-start mechanism that also uses lost datagrams to infer network congestion, and slows the data rate. In effect, these are explicit bandwidth management functions implemented by TCP (the Nagle algorithm, delayed ACKs, and selective ACKs could be considered others), although they rely on implicit notification from the network. [TCP Congestion]

UDP (User Datagram Protocol), on the other hand, does not have any built-in mechanism to detect data loss and slow the send rate. Hence, this functionality must be explicitly enabled by the application itself, or there is nothing to stop a UDP application from sending data, even when network congestion is causing data loss.

Those UDP-based applications that do enable a "back-off" mechanism are called "adaptive applications," and require a feedback loop from receivers to monitor data loss. Real-Time Control Protocol [RTCP] used in combination with the Real-Time Protocol [RTP] provides a standard feedback-loop mechanism. Feedback mechanisms are more difficult to support for multicast distribution, due to the possibility of "implosion" as many receivers respond simultaneously to a sender. Protocols and algorithms to address this challenge are still under research. This is a scalability issue that is also relevant to reliable multicast [ReliableMulticast].

When a receiver feedback is unavailable (e.g. since an "upstream" channel is unavailable), some applications adapt the data-stream itself by sending redundant data using forward error correction (FEC) and/or layered codecs. Redundant data and a back-off mechanism are sometimes employed in tandem to reduce recovery latency. Feedback from QoS protocols can also be used to trigger adaptation, as we describe later.

Active Bandwidth Management

Bandwidth management more commonly refers to active management, and is not currently enabled in most IP networks. This includes a number of algorithms, protocols and APIs, as shown in Table 1.

Some network elements enable "fair queuing" algorithms so a misbehaving application--one that continues to send during times of congestion--won't punish other, better-behaved applications (e.g. TCP applications), or so the average of dropped packets is evenly distributed across flows [Queuing]. Basically, they determine how packets are dropped when congestion occurs in a router (i.e. when a queue is full). CFQ (Class-based Fair Queuing), WFQ (Weighted Fair Queuing), SFQ (Stochastic Fair Queuing) are examples of these algorithms. These are not widely used yet.

RED (Random Early Detection) attempts to avoid congestion rather than reacting to it (and thereby avoid TCP synchronization problems that can result when hosts decrease or increase TCP traffic simultaneously after congestion occurs). It randomly drops packets before queues fill, to keep them from overflowing. Unlike the queue management algorithms mentioned above, it does not require flow-state in the routers.

There have also been proposals to enable explicit congestion notification (ECN)--at least for TCP--but as yet no standards work in this area.

Other research is underway to avoid network congestion by reducing network traffic. For example, "adaptive web-caching" uses (reliable) multicast to dynamically distribute popular web-page caches to local caches that web-browser clients can automatically discover. The intent is to deter web "hot-spots" that can form when "flash-crowds" converge on a website that suddenly becomes popular [WebCache].

Although all of these technologies provide bandwidth management, they are not what many people think of when they use the term. The prevailing meaning of the term refers to Quality of Service (QoS).

What is QoS?

Quality of Service (QoS) is to the ability of a network element (e.g. an application, host or router) to have some level of assurance that its traffic and service requirements can be satisfied. To enable QoS requires the cooperation of all network layers from top-to-bottom, as well as every network element from end-to-end. Any QoS assurances are only as good as the weakest link in the "chain" between sender and receiver.

QoS does not create bandwidth. It isn't possible for the network to give what it doesn't have, so bandwidth availability is a starting point. QoS only manages bandwidth according to application demands and network management settings, and in that regard it cannot provide certainty if it involves sharing. Hence, QoS with a guaranteed service level requires resource allocation to individual data streams.

The bandwidth allocated to an application in a "resource reservation," is then unavailable for use by "best-effort" applications. Considering that bandwidth is a finite resource, a priority for QoS designers has been to ensure that best-effort traffic is protected, after reservations are made. QoS-enabled (high-priority) applications must not disable the mundane (low-priority) Internet applications. The worst case should be that low-priority applications simply have a lesser (slower) service, but still function.

There are essentially two types of QoS available:

- ***Resource reservation*** (integrated services): network resources are apportioned according to an application's QoS request, and subject to bandwidth management policy. RSVP provides the mechanisms to do this.

- ***Prioritization*** (differentiated services): network traffic is classified and apportioned network resources according to bandwidth management policy criteria. To enable QoS, classifications give preferential treatment to applications identified as having more demanding requirements. DiffServ provides this service.

These QoS protocols and algorithms are not competitive or mutually exclusive, but on the contrary, they are complementary. As a result, they are designed for use in combination to accommodate the varying operational requirements in different network contexts. Figure 2 illustrates a complete architecture in which they work together to provide end-to-end QoS across multiple service providers. Notice that this includes a number of other service components that we describe later.

QoS	***Net***	***App***	***Description***
most	X		Provisioned resources end-to-end (e.g. private, low-traffic network)
	X	X	RSVP (Resource reSerVation Protocol) Guaranteed Service (provides feedback to application)
	X	X	RSVP Controlled Load Service (provides feedback to application)
	X		Multi-Protocol Label Switching (MPLS)
	X	X	Diffserv (Differentiated Services) applied at network core ingress appropriate to RSVP reservation service level for that flow.
	X	X	Diffserv applied on per-flow basis by source application (IP stack)
	X		Diffserv applied at network core ingress
	X		Best effort service with explicit congestion notification (ECN)
	X		Fair queuing applied by network elements (e.g. CFQ, WFQ, RED)
least			Best effort service (implicit congestion notification for TCP only)

Table 3. Shows the different bandwidth management algorithms and protocols, their relative QoS levels, and whether they are activated by network elements or applications, or both.

RSVP - Resource Reservation

RSVP is a reservation setup and control protocol that enables the integrated services designed to provide the closest thing to circuit-emulation on IP networks. RSVP is the most complex of all the QoS technologies, for applications (hosts) and for network elements (routers and switches). As a result, it also represents the biggest departure from the tried-and-true "best-effort" IP network, which creates some cause for concern.

Here is a simplified overview of how the protocol works:

- Senders characterize outgoing traffic in terms of the upper and lower bounds of bandwidth, delay, and jitter. RSVP sends a PATH message that contains this traffic specification (TSpec) information to the (unicast or multicast) destination address. Each RSVP-enabled router along the downstream route establishes a "path-state" that includes the previous source address of the PATH message (i.e. the next hop "upstream" towards the sender).

- To make a resource reservation, receivers send a RESV (reservation request) message "upstream" to the (local) source of the PATH message. In addition to the TSpec, the RESV message includes the QoS level required (controlled load or guaranteed) in an RSpec, and characterizes the packets for which the reservation is being made (e.g. the transport protocol and port number), called the "filter spec." Together, the RSpec and filter-spec represent "flow-descriptor" that routers use to identify reservations.

- When an RSVP router receives an RESV message, it uses the admission control process to authenticate the request and allocate the necessary resources. If the request cannot be satisfied (due to lack of resources or authorization failure), the router returns an error back to the receiver. If accepted, the router sends the RESV upstream to the next router.

- When the last router receives the RESV and accepts the request, it sends a confirmation message back to the receiver (note: the "last router" is either closest to the sender or at a reservation merge point for multicast flows).

- There is an explicit tear-down process for a reservation when sender or receiver ends an RSVP session.

Here are some salient characteristics of RSVP support:

- Reservations in each router are "soft," which means they need to be refreshed periodically by the receiver(s).

- RSVP is not a transport, but a network (control) protocol. As such, it does not carry data, but works in parallel with TCP or UDP data "flows."

- Applications require APIs to specify the flow requirements, initiate the reservation request, and receive notification of reservation success or failure after the initial request and throughout a session. To be useful, these APIs also need to include RSVP error information to describe a failure.

- Multicast reservations are "merged" at traffic replication points on their way upstream, which involves complex algorithms.

- Although RSVP traffic can traverse non-RSVP routers, this creates a "weak-link" in the QoS chain where the service falls-back to best effort (i.e. all bets are off).

DiffServ - Prioritization

Differentiated Services [DiffServ] provides a simple and coarse method of classifying services of various applications. There are currently two per hop behaviors (PHBs) defined in draft specifications that in effect represent two service levels (traffic classes)--assured and premium. These PHBs are applied to traffic at a network ingress point (network border entry) according to pre-determined policy criteria. The traffic may be marked at this point, and routed according to the marking, then unmarked at the network egress (network border exit).

DiffServ assumes the existence of a service level agreement (SLA) between networks that share a border. The SLA establishes the policy criteria, and defines the traffic profile. It is expected that traffic will be policed and smoothed at egress points, and any traffic "out of profile" (i.e. above the upper-bounds of bandwidth usage stated in the SLA) at an ingress point have no guarantees (or may incur extra costs, according to the SLA). The policy criteria used can include time of day, source and destination addresses, transport, and/or port numbers (i.e. application IDs). Basically, any context or traffic content (including headers or data) can be used to apply policy.

When applied, the protocol mechanism that the service uses are bit patterns in the "DS-byte," which for IPv4 is Type-of-Service (TOS) octet and for IPv6 is the Traffic Class octet. There is some debate as to whether the original IPv4 TOS bit values as defined by RFC 1349 will be preserved (most likely they will be, since they are currently used within some enterprises and ISPs).

The simplicity of DiffServ to prioritize traffic belies its flexibility and power. When DiffServ uses RSVP parameters or specific application types to identify and classify CBR traffic, it will be possible to establish well-defined aggregate flows that may be directed to fixed bandwidth pipes. As a result, you could share resources efficiently and still provide guaranteed service.

MPLS - Label Switching

Multi-Protocol Label Switching (MPLS) is similar to DiffServ in some respects, as it also marks traffic at ingress boundaries in a network, and unmarks at egress points. But unlike DiffServ, which uses the marking to determine priority within a router, the markings are designed to determine the next router to go to. In other words, MPLS affects routing services and can be used to establish "fixed bandwidth pipe" analogous to an ATM or Frame Relay virtual circuit.

As described in [MPLS-RSVP], there's a proposal to add new objects to RSVP to enable allocation of network resources for use by MPLS "pipes." These pipes are virtual paths established through a network when routers use MPLS tags determine where to forward traffic, and are illustrated in Figure 2.

Layer 3 Switching

Layer-3 switches represent a new breed of network element product that enable much of an IP router's functionality with the simplicity, efficiency, and lower-cost of a network switch. They use IP network layer information, as well as Layer 4 (TCP or

UDP port numbers) or application attributes, to apply policy for traffic classification, filtering and forwarding. Like standard routers, Layer-3 switches do not mark/unmark traffic.

Active Networks

On the high-end of the spectrum of active management technologies are Active Networks [ActiveNets]. This refers to providing a programmable interface in network elements. In a sense, policy application/enforcement enables a way to program a network element that affects traffic control dynamically. However, true programmability enables dynamic definition and deployment of new protocols in addition to per-hop traffic manipulation. This is a promising area of research, but still primarily limited to the labs.

End-to-End QoS Model

Figure 2 shows a complete picture of how some of these QoS technologies can work together to provide "end-to-end QoS". [e2e-QoS]. Aside from the bandwidth broker--which is still a new concept at this point in time--this represents the model under development within the IETF community.

RSVP provisions resources for network traffic, whereas DiffServ simply marks and prioritizes traffic. RSVP is more complex and demanding than DiffServ in terms of router requirements, so can negatively impact backbone routers. This is why the "best common practice" says to limit RSVP's use on the backbone [AppStatement], and

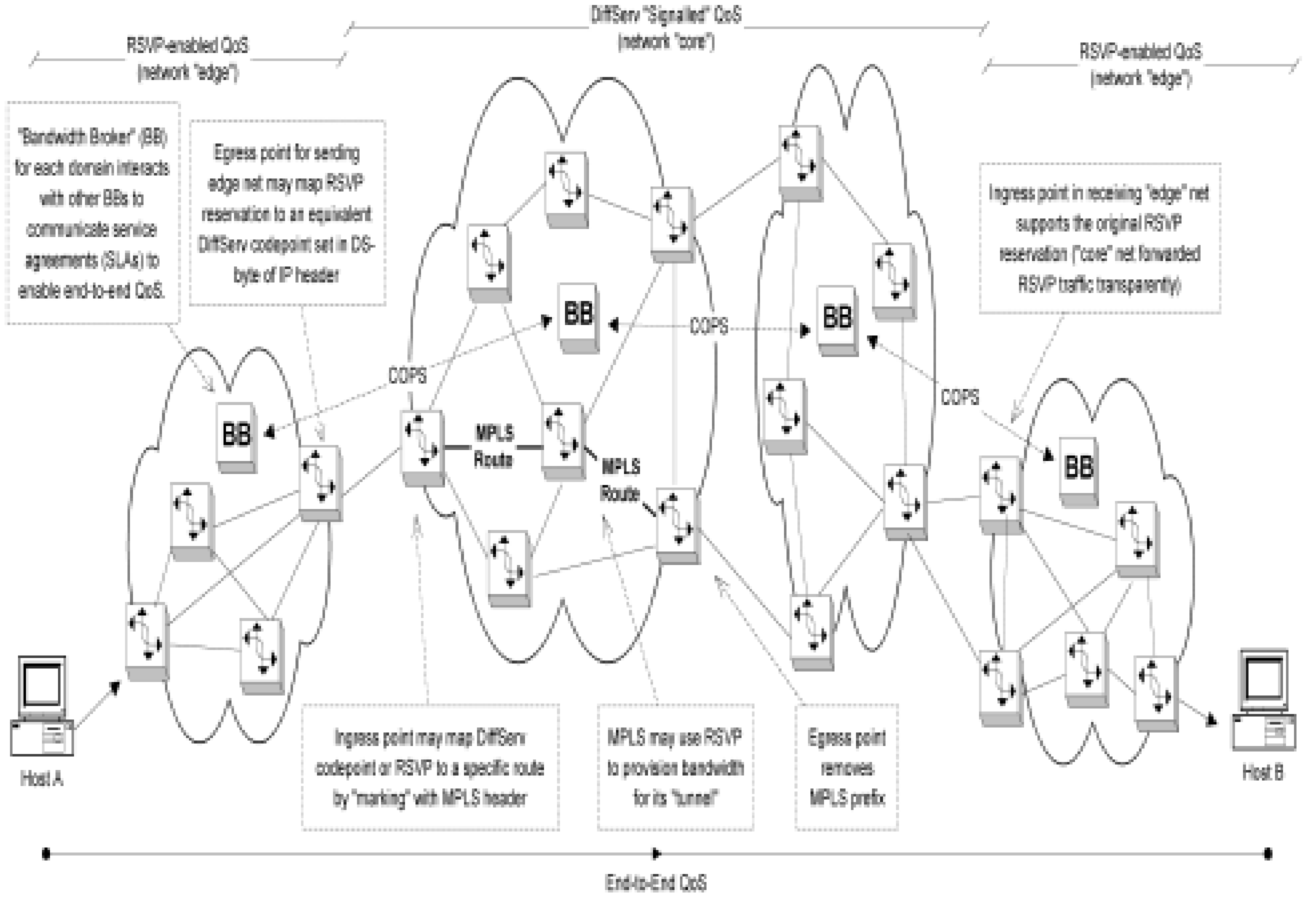

DiffServ "Signalled" QoS
(network "core")
RSVP-enabled QoS
(network "edge")
RSVP-enabled QoS
(network "edge")
"Bandwidth Broker" (BB) for each domain interacts with other BBs to communicate service agreements (SLAs) to enable end-to-end QoS.
Egress point for sending edge net may map RSVP reservation to an equivalent DiffServ codepoint set in DS-byte of IP header
Ingress point in receiving "edge" net supports the original RSVP reservation ("core" net forwarded RSVP traffic transparently)
BB
BB
BB
BB
COPS
COPS
COPS
MPLS Route
MPLS Route
Host A
Host B
Ingress point may map DiffServ codepoint or RSVP to a specific route by "marking" with MPLS header
MPLS may use RSVP to provision bandwidth for its "tunnel"
Egress point removes MPLS prefix
End-to-End QoS

Figure 2: Illustrates the possible use of different QoS technologies under development--RSVP, DiffServ, MPLS, COPS and Bandwidth Brokers--working cooperatively in various stratagies to enable end-to-end QoS*

why DiffServ *can* exist there.

DiffServ is a perfect compliment to RSVP as the combination can enable end-to-end quality of service (QoS). End hosts may use RSVP requests with high granularity (e.g. bandwidth, jitter threshold, etc.). Border routers at backbone ingress points can then map those RSVP "reservations" to to a class of service indicated by a DS-byte (or source host may set the DS-byte accordingly also). At the backbone egress point, the RSVP provisioning may be honored again, to the final destination. Ingress points essentially do traffic conditioning on a customer basis to assure that service level agreements (SLAs) are satisfied.

The architecture represented in Figure 2--RSVP at the "edges" of the network, and DiffServ in the "core"-- has momentum and support. Work within the IETF DiffServ work group is progressing quickly, although initial tests have shown mixed results.

Top-to-Bottom QoS

As mentioned earlier, QoS guarantees are only as good as their weakest link. Since many networks use switched Ethernet, and Ethernet is a shared media, and shared bandwidth creates a weak link, we should mention the work to define QoS for Ethernet.

The IEEE 802.1p and 802.1q standards define how Ethernet switches can classify traffic in order to expedite delivery of time critical traffic. The IETF Integrated Services over Specific Link Layers [issll] working group is chartered to define the mapping between upper (network) layer QoS protocols and services, with those of the lower (data-link) layer. This work is still underway at the time of this writing.

Business Opportunities

By enabling QoS--which essentially allows one user to get a better service than another--we create an incentive to steal. As a result, QoS requires policy enforcement that will require a policy management infrastructure.

However, we cannot enforce policy unless we can establish the identities of network users and thereby assign a level of trust. This implies a need for an authentication infrastructure.

And of course, since QoS provides added value, it has additional worth. Hence, QoS also implies a need for accounting and billing infrastructure.

These three support services--***Policy Management, Authentication,*** and ***Accounting/Billing***--are essential to the success of QoS. All of them represent technical challenges that are being addressed, but more importantly, they represent significant new business opportunities that will provide further incentive to QoS deployment.

Another very important scaling consideration that applies to all three of these services is managing peering arrangements between various Internet Service Providers (ISPs). Before we can enable end-to-end QoS, bilateral agreements must be in place so ISPs sharing QoS responsibilities for common flows can share the necessary policy, authentication and accounting/billing information.

Policy Management

The Common Open Policy Service (COPS) protocol is emerging as a viable solution for distributed policy management [PolicyFramework]. Initially, COPS will be used within a domain, for router policy enforcement points (PEPs) to retrieve policy from the policy distribution points (PDPs). As pictured in Figure 2, COPS may be also be used between Bandwidth Brokers (BBs)--which essentially act as PDPs--for dynamic inter-domain policy exchange.

One possibility is that Bandwidth Brokers could be third parties that manage SLAs for various ISPs and enterprises.

Authentication Infrastructure

Work in the IETF public key infrastructure (PKIX) work group was recently finalized and since Internet commerce via SSL-enabled "web-shopping" has been using "certificates" for a number of years already, the technologies have had a chance to mature. There is still some work to be done, and things may yet change as wrinkles are ironed-out, but the certificate infrastructure to provide credentials for user/server/enterprise authentication and thereby enable trust on the Internet is good and getting better [PKIX-Roadmap]. Legislation has been passed in a number of states to legally validate the technologies.

The foundation of the trust model that the PKIX establishes are "certificate authorities," which issue and revoke certificates with different levels of verification and associated liability. There are many analogies, such as notary public and bonding. These certificate authorities may provide tremendous opportunities for third party branding. The model is similar in many ways to the one implemented by Visa and Mastercard, where their credit cards represent certificates that enable transactions between the customers and banks.

Accounting and Billing

The work in this area has only recently started in the IETF. IETF 42 had a BOF for discussion of forming a work group to address the standards needs for authentication, authorization and accounting (AAA). As with so much in the area of Bandwidth Management, this is closely related to support of IP Telephony and may borrow from that industry.

It would seem that a Bandwidth Broker third party could manage the billing and accounting for a number of ISPs, in addition to their SLAs. This concept is not unlike the current arrangements in the de-regulated telecommunications industry. Third parties broker on behalf of phone customers to find the best carriers and RBOCs

(Regional Bell Operating Companies) for local and long-distance services at any point in time. The Iridium global satellite phone system scheduled to go on-line in November '98 reportedly has the most complex billing systems developed to date to deal with the many phone systems throughout the world.

Conclusion

The global Internet has changed everything, and as a result the world is converging on the Internet Protocol for its networking needs. But in the process, the weaknesses in IP's original design have been exposed. As a result, the Internet needs to change to accommodate new application demands. Bandwidth is needed, but it isn't enough. In addition, the Internet needs bandwidth management. Specifically, it needs some "intelligence."

Until now, IP has provided a "best-effort" service in which network resources are shared equitably. Adding QoS raises significant concerns, since it enables differentiated services that represent a significant departure from the fundamental and simple design principles that made the Internet a success. These new QoS services also require support services of their own, all of which have not been completely-defined, let alone deployed at this point in time.

On the other hand, these new services will create tremendous new business opportunities the prospect of which generate significant incentive for their support. Hence, they are an inevitable evolution for the Internet, and through the pain of change, we will all gain.

The final result will be a network that can all but eliminate the constraints of physical distance, insofar as it can satisfy human senses. In the short term, multi-media conferences will be possible. The ultimate goal, however, will be to transport enough information with a minimum of delay so that we can experience a virtual presence of anywhere in the globe, with sensory feedback and real-time control. It's pure science fiction at this point, but then again, so was the idea of a ubiquitous and seamless global Internet a mere 20 years ago.

References

[ActiveNets] IEEE Communications Magazine, *Programmable Networks*, October 1998 (Vol. 36 No. 10), http://www.comsoc.org. Also see http://www.ieee-pin.org

[AppStatement] A. Mankin, et al, *Resource ReSerVation Protocol (RSVP) Version 1 Applicability Statement - Some Guidelines on Deployment*, RFC 2208, September 1997

[Controlled] J. Wroclawski, *Specification of the Controlled-Load Network Element Service*, RFC 2211, Sept 1997

[DiffServ] IETF "Differentiated Services" working group. See http://www.ietf.org/html-charters/diffserv-charter.html

[DNS SRV RR] A. Gulbrandeen, P. Vixie, *A DNS RR for specifying the location of services (DNS SRV)*, RFC 2052, October 1996

[e2e] J. Saltzer, D. Reed, D. Clark, *End to End Arguments in System Design*, ACM Transactions in Computer Systems, November 1984. See http://www.reed.com/Papers/EndtoEnd.html

[e2e-QoS] Y.Bernet, R.Yavatkar, P.Ford, F.Baker, L.Zhang, K.Nichols, M.Speer, *A Framework for Use of RSVP with Diff-serv Networks*, <draft-ietf-diffserv-rsvp-00.txt>, Work in Progress

[ECN] "Explicit Congestion Notification," see http://www-nrg.ee.lbl.gov/floyd/ecn.html

[Gecsei] Jan Gecsei, *Adaptation in Distributed Multimedia Systems*, IEEE MultiMedia, April/June 1997 Vol. 4, No. 2 http://computer.org/multimedia/mu1997/u2058abs.htm

[Grove] Andrew S. Grove, chairman and co-founder of Intel, as quoted by George Gilder in *The Bandwidth Tidal Wave*, in his Telecosm series. See http://www.forbes.com/asap/gilder/

[Guaranteed] S. Shenker, C. Partridge, R. Guerin, *Specification of Guaranteed Quality of Service*, RFC 2212, Sept 1997

[IETF] The Internet Engineering Task Force is a loose confederacy of volunteers from the network industry and academia that uses "running code and rough consensus" to establish protocol standards for the Internet.

[IntServ] IETF "Integrated Services" working group. See http://www.ietf.org/html-charters/intserv-charter.html

[issll] Integrated Services over Specific Link Layers, see http://www.ietf.org/html.charters/issll-charter.html

[Maufer] Maufer, T., Deploying IP Multicast in the Enterprise, Prentice-Hall, December 1997, ISBN 0-13-8976872

[Moore] In 1979 Intel's chairman and co-founder, Gordon Moore, observed and projected that computing power--based on the density of transistors in a single chip--doubles every 18 months. To a surprising extent, this rule-of-thumb for computing power developments has continued to hold true to this day.

[MPLS] IETF "Multiprotocol Label Switching" working group. See http://www.ietf.org/html-charters/mpls-charter.html

[MPLS-RSVP] D.Awduche, D.Gan, T.Li, G.Swallow, V. Srinivasan , "Extensions to RSVP for Traffic Engineering", <draft-swallow-mpls-rsvp-trafeng-00.txt>, Work in Progress

[Paradigm] C. Lefelhocz, B. Lyles, S. Shenker, L. Zhang, *Congestion Control for Best-Effort Service: Why We Need a New Paradigm*, IEEE Network, Jan/Feb 1996, Volume 10, Number 1.

[PKIX Roadmap] A. Arsenault, *Internet X.509 Public Key Infrastructure PKIX Roadmap*, <http://www.imc.org/ietf-pkix/mail-archive/1949.html> or <ftp://ftp.ietf.org/internet-draft/draft-ietf-pkix-roadmap-00.txt>, Work in Progress

[PolicyFramework]	R. Yavatkar, D. Pendarakis, R. Guerin, *A Framework for Policy-based Admission Control*, <draft-ietf-rap-framework-00.txt>, Work in Progress
[PolicyLanguage]	J.Strassner, Ed Ellesson, *Terminology for describing network policy and services*, <draft-strassner-policy-terms-00.txt>, Work in Progress
[Queuing]	Nortel/Bay Networks, *IP QoS - A Bold New Network*, white paper, see http://www.nortel.com/home/images/IP_QOS_WP.pdf or http://www.nortel-bay.com/english/solutions/qos_wp.pdf
[ReliableMulticast]	IP Multicast Initiative (IPMI), *State-of-the-Art: Reliable IP Multicast*, March 1998. See http://www.ipmulticast.com
[Schmidt]	Schmidt, A.G., Minoli, D., Multiprotocol over ATM: Building State of the Art ATM Intranets, (c) 1998 Manning Publications Co., ISBN 1-884777-42-2
[ServiceSpec]	S. Shenker, J. Wroclawski, *Network Element Service Specification Template*, RFC 2216, Sept 1997
[SLAs]	N. Brownlee, *SRL: A Language for Describing Traffic Flows and Specifying Actions for Flow Groups*, <draft-ietf-rtfm-ruleset-language-02.txt>, Work in Progress
[SrcQuench]	[Stevens] notes that although RFC 1009 says a router must generate ICMP source quench messages when it runs out of buffers, RFC 1716--the updated Router Requirements--says a router should not generate source quench (since they consume network bandwidth and are considered an ineffective and unfair fix for congestion).
[SrvLoc]	J. Veizades, E. Guttman, C. Perkins, S. Kaplan, *Service Location Protocol*, RFC 2165, June 1997
[Stevens]	W. Richard Stevens, TCP/IP Illustrated, Volume I, (c)1994, Addison-Wesley, ISBN 0-201-63346-9,
[TCP Congestion]	W. Stevens, *TCP Slow Start, Congestion Avoidance, Fast Retransmit, and Fast Recovery Algorithms*, RFC 2001, January 1997
[TOS]	Almquist, P. *Type of Service in the Internet Protocol Suite*, July 1992, RFC 1349
[WebCache]	See http://www-nrg.ee.lbl.gov/floyd/web.html

Glossary

Admission Control	Policy decision applied initially to QoS requests (not to be confused with "policing," which occurs after a request is accepted and data is flowing). Admission Control is closely tied to accounting, and relies on source authentication.
ALTQ	Alternative Queuing: A public domain FreeBSD implementation of CBQ, RED and WFQ. See http://www.csl.sony.co.jp/person/kjc/papers/usenix98/altq.html
ATM	Asynchronous Transfer Mode
CAT5	"Category-5" unshielded twisted pair for high-speed data traffic over short distances

CBR	Constant Bit Rate
CFQ	Class-based Fair Queuing: per-class packet scheduling. See http://www-nrg.ee.lbl.gov/floyd/cbq.html
CIR	Committed Information Rate
Class	An abstraction that can be determined by different policy criteria such as IP packet header content (e.g., source or destination addresses or port numbers or transport protocol), or time of day, ingress point, etc. The definition of a class can differ at different locations on the network.
Controlled Load	Tightly approximates best-effort service under unloaded conditions
CoS	Class of Service: differentiated services based on traffic types (e.g. high-priority two-way isochronous traffic or constant bit rate services, versus low-priority file transfer or email delivery, with interactive web-browsing somewhere in between)
CSMA/CD	Carrier Sense Multiple Access/Collision Detect - the media access mechanism used by Ethernet to 1) look for carrier 2) begin transmitting when none found 3) detect collisions, and 4) back-off for a random interval before trying again
ECN	Explicit Congestion Notification, see http://www-nrg.ee.lbl.gov/floyd/ecn.html
EF	Expedited Forwarding
FEC	Forward Error Correction: Sending redundant data so that if data loss occurs, data recovery is possible without retransmission
FIFOQ	First-In First-Out Queuing: simple tail-drop FIFO queue used in (best effort) IP service.
Flow	"A flow is a set of packets traversing a network element, all of which are covered by the same request for control of quality of service" [ServiceSpec]. It is considered something between a pure virtual circuit and pure datagram.
Guaranteed Service	Delay-bounded service with no queuing loss
Jitter	Variations in delay
LAN	Local-Area Network
MAN	Metropolitan Area Network
Network Element	Any component of an internetwork which directly handles data packets, and thus is potentially capable of exercising QoS control over data flowing through it.
PHB	Per Hop Behavior
Policing	Packet-by-packet monitoring function at a network border (ingress point) that ensures a host (or peer, aggregate, whatever) does not violate its promised traffic characteristics
QoS	Quality of Service: integrated services that satisfy delay and bandwidth parameters

RED	Random Early Detection: for congestion avoidance and notification. Unlike CBQ, WFQ, SFQ, RED does not require flow-state in routers. Randomly drops datagrams when congestion detected (e.g. queue overflows). See http://www-nrg.ee.lbl.gov/floyd/red.html
RSVP	Resource ReSerVation Protocol, see http://www.isi.edu/rsvp/
SFQ	Stochastic Fair Queuing: hash function used to map flow to one of set of queues
SLA	Service Level Agreement: Contract between Service provider and their customer that defines provider responsibilities in terms of network levels (throughput, loss rate, delays and jitter) and times of availability, method of measurement, consequences if service levels aren't met or the defined traffic levels are exceeded by the customer, and all costs involved.
SMDS	Switched Multi-megabit Data Service
SONET	Synchronous Optical Network
UTP	Unshielded Twisted Pair
VBR	Variable Bit Rate
WAN	Wide-Area Network
WDM	Wavelength Division Multiplexing
WFQ	Weighted Fair Queuing: per-flow packet scheduling in network elements
xDSL	Digital Subscriber Line, which represents a number of different--but similar--technologies. The 'x' is replaced according to the type (A: Asynchrnous, S: Single-Line, V: Very High-Speed)