Quality of Service—Fact or Fiction?

by Geoff Huston, Telstra

Much has been written about the potential of *Quality of Service* (QoS) and the Internet. However, much of the material is strong on promise, but falls short in critical analysis. In an effort to balance the picture, we present here a brief status report on the QoS effort, exposing some of the weaknesses in the current QoS architectures.

The QoS Service

The default service offering associated with the Internet is a *best-effort* service, where the network treats all traffic in exactly the same way. There is no consistent service outcome from the Internet best-effort service model. When the load level is low, the network delivers a high-quality service. The best-effort Internet does not deny entry to traffic, so as the load levels increase, the network congestion levels increase, and service-quality levels decline uniformly. This decline in service is experienced by all traffic passing through a congestion point, and is not limited to the most recently admitted traffic flows. For many applications, this best-effort response is perfectly acceptable. When network capacity is available, the application can make use of the resource, whereas when the level of contention for network bandwidth is high, each application will experience similar levels of congestion. A best-effort network service is a good match to opportunistic applications that can vary their data transfer rate in response to signaled network load.

The objective of various Internet QoS efforts is to augment this service with a number of selectable service responses. These service responses may be different from the best-effort service by some form of superior service response, such as lower delay, lower jitter, or greater bandwidth. These responses are relative, where the service outcome is claimed to be no worse than best effort at any time, and superior to best-effort under congestion load. Alternatively, QoS service responses may be distinguished by providing a consistent, and therefore predictable, service response that is unaffected by network congestion levels. These are quantitative service responses, where the characteristics of the service can be measured against a constant outcome. A quantitative service many be one that constrains jitter to a maximum level, or one that makes a certain bandwidth available, within parameters of bounded jitter, similar to a conventional leased line. Such constant-rate services may be superior to best-effort services when the network is under load, but they may also offer inferior service when the network is under negligible load. The essential attribute of these services is one of consistency.

Why is there a need for relative or consistent service profiles within the Internet? The underlying reasons for introducing QoS into the Internet appear to be threefold: First is the desire to provide high-quality support for IP voice and video services, second is the desire to manage the service response provided to low-speed access devices, such as Internet mobile wireless devices, and third is the desire to provide a differentiated Internet access service, providing a network client with a range of service-quality levels at a range of prices.

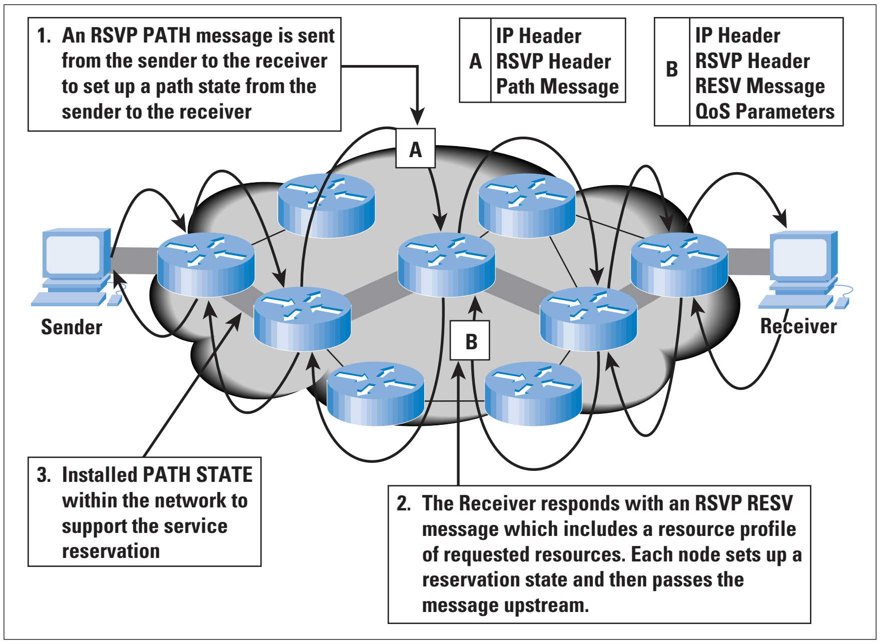
Obviously this is a broad agenda, where there are requirements to extend specific network services to applications, requirements to adapt network services to particular transmission characteristics, and requirements to manage network resources to achieve particular response characteristics for an aggregated collection of traffic.

Approaches to QoS

The relevant efforts within the *Internet Engineering Task Force* (IETF) have been addressing standards for QoS mechanisms within the network.

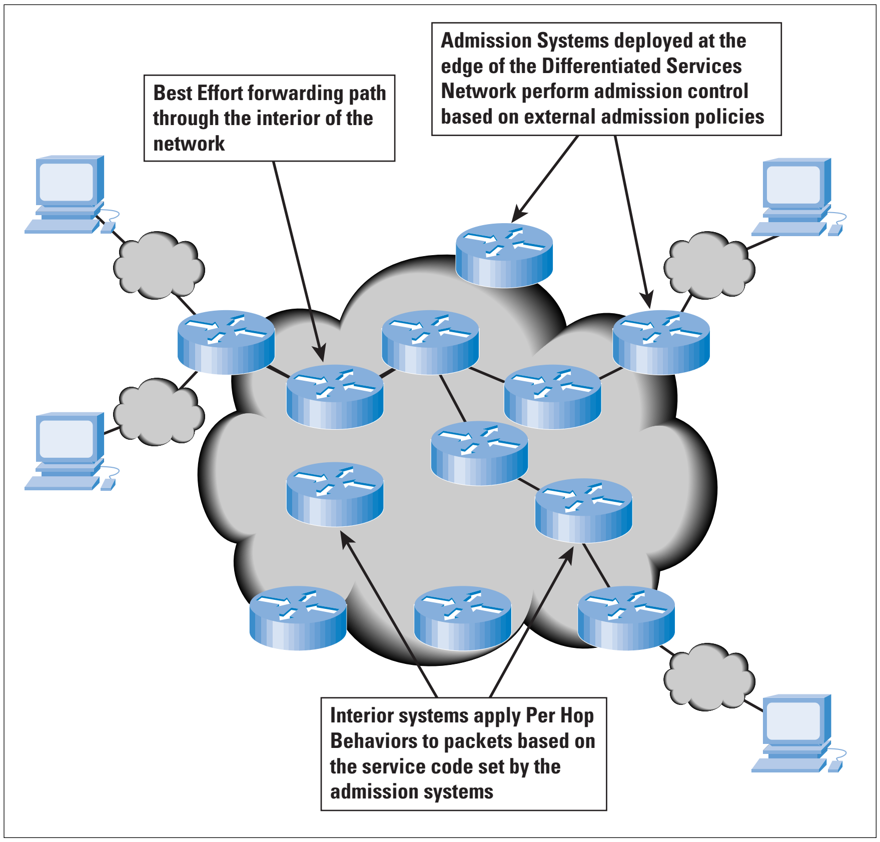
The initial approach to QoS was that of the *Integrated Services* architecture. This approach focuses on the application as the trigger for QoS. Here, the application first signals its service requirements to the network in the form of a reservation, and the network responds to this request. The application proceeds only if the network has indicated that it is able to carry the additional load at the requested service level by committing to the reservation. The reservation remains in force until the application explicitly requests termination of the reservation, or the network signals to the application that it is unable to continue the reservation. The essential feature of this model is the "all-or-nothing" nature of the service model. Either the network commits to the reservation, in which case the application does not have to monitor the level of network response to the service, or the network indicates that it cannot meet the reservation. This approach imposes per-application state within the network, and for large-scale networks, such as the global Internet itself, this approach alone does not appear to be viable (see Figure 1).

Figure 1: The Integrated Services QoS Architecture


1. An RSVP PATH message is sent from the sender to the receiver to set up a path state from the sender to the receiver
A
IP Header
RSVP Header
Path Message
B
IP Header
RSVP Header
RESV Message
QoS Parameters
A
Sender
B
Receiver
3. Installed PATH STATE within the network to support the service reservation
2. The Receiver responds with an RSVP RESV message which includes a resource profile of requested resources. Each node sets up a reservation state and then passes the message upstream.

The subsequent approach to QoS mechanisms has been to look at the core of the network, and examine those mechanisms that can provide differentiated service outcomes with appropriate scaling properties. This approach, the *Differentiated Services* architecture, includes dropping the concept of a per-application path state across the network using instead the concept of aggregated service mechanisms. Within the aggregated service model, the network provides a smaller number of different service classes and aggregates similar service demands from a set of applications into a single service class. Aggregated services are typically seen as an entry filter, where on entry to the network each packet is classified into a particular service profile. This classification is carried within the IP packet header, using 6 bits from the deprecated IP *Type of Service* (TOS) header to carry the service coding. The network then uses this service code in the packet header to treat this packet identically to all other packets within the same service code. While this approach does possess the ability to scale across the entire Internet, there are numerous unresolved issues relating to the quality signaling between individual applications and the network. The aggregated service model does not allow an individual application to sense if it is receiving the necessary service response from the network (see Figure 2).

Figure 2: The Differentiated Services QoS Architecture


Best Effort forwarding path through the interior of the network
Admission Systems deployed at the edge of the Differentiated Services Network perform admission control based on external admission policies
Interior systems apply Per Hop Behaviors to packets based on the service code set by the admission systems

QoS Deployment

Neither approach alone is adequate to meet the QoS requirements. The Integrated Services approach alone imposes an excessive load in the core of large networks through the imposition of a per-application path state. The Differentiated Services approach does provide superior scaling properties through the use of aggregated service elements, but includes no concept of control signaling to inform the traffic conditioning elements of the current state of the network, or the current per-application requirements.

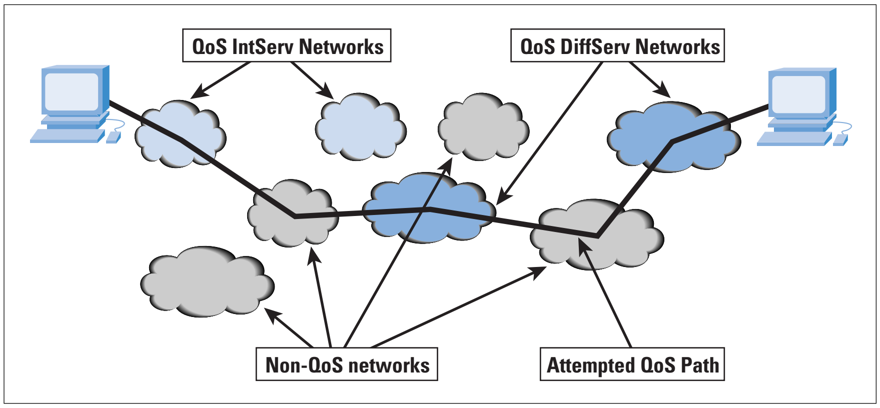
The underlying question then becomes: Is a combination of these two approaches sufficient to allow QoS to be widely deployed on the Internet?

At this stage the response does appear to be a "No." Perhaps this strong negative response should be further qualified. The existing tools are insufficient to support widespread use of QoS-based services on the multiprovider public Internet. The qualification is that within the enterprise network environment there are much stronger drivers for QoS mechanisms and much greater levels of administrative control over the overall network architecture, while within the multiprovider public Internet, these drivers are not apparent. The enterprise approach may also have some parallels within a single IP carrier's network, or even across some forms of bilateral agreements between carriers. However, such approaches are not anticipated to be a widespread feature of the public Internet service environment.

Let's look more closely at the public Internet and QoS to see why there is a mismatch between the two. The major stumbling blocks in attempting to address how QoS could be deployed in the public Internet are both engineering and economic in nature.

From an engineering perspective, we need to remember that in order to actually deliver any reasonable assurance of a quality-differentiated service, the service-quality mechanism chosen must be deployed across all networks along the end-to-end paths of the quality-service traffic. In a heterogeneous multiprovider environment such as the public Internet, this outcome is very unlikely. Within the tens of thousands of component service providers that make up the global Internet, such uniformity of action is highly improbable. The IPv6 transition structure correctly identifies the first step as isolated "islands" of IPv6 functionality, interconnected by some form of IPv6 "bridges." While the potential scenario of initial QoS deployment may be similar, in terms of isolated islands of deployment of QoS services, there is a much stricter requirement for the "bridges" across the non-QoS-aware parts of the network; namely, that they do not distort the service outcomes. In effect, this scenario requires a QoS response from a non-QoS system (see Figure 3).

Figure 3: Attempted End-to-End QoS across the Public Internet



The engineering issues are deeper than simply the considerations of transition within a potential deployment scenario. The issues include:

- The need for QoS-enabled applications that can predict their service requirements in advance, and be able to signal these requirements into the network.
- In the case of the differentiated service approach of admission controls, there is a requirement for the interior of the network to be able to signal current load conditions to the network admission systems. This system also requires that the admission control points be able to use admission-decision support systems in order to include consideration of the service load, the current network load, and the policy parameters of the network that may allow some level of preemption of various admission decisions in order to meet high-priority service requirements.
- The signaling and negotiation aspect of QoS extends into the interdomain space, where two or more service providers need to negotiate mutually acceptable service profiles, and associated service access. This extends beyond the addition of bilateral agreements and encompasses the requirement to add QoS attributes to interdomain routing protocols. The tools and operating techniques required to support this functionality remain poorly defined.
- Measurement of service performance remains an area in which existing measurement tools are lacking. While it is possible to instrument every active device within a network into a network management system, such an element-by-element view does not readily translate to the end-to-end view of application service performance.

From an economic perspective, we must remember that no current Internet retail tariff includes a concept of end-to-end tariffed transactions. All tariffs are access based, because application transactions are not readily visible to the Internet network. In addition, no technically stable or financially stable structure of interprovider interconnection financial settlements exists today. The financial model of the Internet from an economic viewpoint is very polarized, with only customer and zero-dollar peer arrangements dominating the interprovider space. However, end-to-end QoS transactions demand a different economic model.

The initiator of the end-to-end QoS transaction has the discretion of choosing whether to request an end-to-end service profile. If such a profile is requested, the initiator should pay the initiating provider a retail tariff to cover the entire end-to-end cost of the transaction, and the initiating provider must then indicate a willingness to financially settle with transit peer networks in order for these transit peers to devote additional resources to service the traffic associated with this transaction, and so forth through the entire path of transit providers. The arbitrary nature of the Internet transits, the dynamic nature of routing, and the lack of transaction setups in any scalable form of QoS mechanisms make this entire scenario highly improbable within our current understanding of interprovider policy-management mechanisms.

The relatively loosely coordinated structure of the public Internet will have to change from the state we have today if we want to use QoS-based services. The changes include:

- A common selection of a set of QoS mechanisms to deploy,
- Ubiquitous deployment of these mechanisms across both service provider and client networks,
- The adoption of a uniform set of retail tariffs for QoS services,
- The definition and common acceptance of multi-party QoS-related financial settlements that support fair and equitable cost distribution among multiple providers, and
- The definition of commonly accepted service performance metrics and related measurement methodologies to allow end-to-end and network-by-network service outcomes to be objectively assessed.

This is a significant agenda for the industry at large to undertake, and more so in an environment that features diversity and vigorous competition between various public Internet service providers.

An additional factor is also working against QoS deployment in the public Internet space. The increasing availability of very-high-speed transmission systems is bringing network carriage capacity down to the level of an abundant commodity across large parts of the Internet world. As the unit costs of network capacity decline in the face of increasing levels of availability of transmission systems, the market niche that QoS could occupy in managing a scarce resource is shrinking. The driver for QoS deployment is not that the best-effort service is not good enough. The problem that QoS is attempting to address is one of allocation of network capacity at those points in time when the network is under heavy load, or, in other words, taking on the task of allocating capacity when there is not enough network capacity to meet every demand. When a network is under load, the QoS response is to place additional control functionality in both applications and in the network to manage this allocation function. Obviously such an activity imposes additional costs on the network operators and the network client. Such additional costs have not created any additional network capacity.

The total sum of demand remains in excess of capacity after the deployment of QoS mechanisms. The alternative approach is to incur additional cost by augmenting the capacity of the network. This approach minimizes the impact of load on the network causing disruption to individual transactions. Again this approach imposes additional costs onto the network, but in an environment of abundant transmission capacity, it may often be the more cost-effective approach.

Where does this leave QoS and the public Internet? There is no doubt that QoS is a very stimulating area of research, with much to offer the enterprise network environment, but in asking for QoS to be deployed within the existing incarnation of the public multiprovider Internet, we may be simply asking for too much at this point in time. More effort is required to turn a QoS Internet into a reliable production platform.

Further Reading

[1] Huston, G., *Internet Performance Survival Guide: QoS Strategies for Multiservice Networks,* ISBN 0471-378089, John Wiley & Sons, January 2000.

A detailed examination of Internet Quality of Service technologies and their potential application within the Internet.

[2] Kilkki, K., *Differentiated Services for the Internet,* ISBN 1578701325, Macmillan Technical Publishing, June 1999.

An in-depth look at the Differentiated Services architecture and its use in enabling networks to handle traffic classes in a specific manner.

[3] Durham, D., and Yavatar, R., *Inside the Internet's Resource Reservation Protocol: Foundations for Quality of Service,* ISBN 0471322148, John Wiley & Sons, April 1999.

At the core of the Integrated Services architecture is a signaling protocol to undertake service reservations. The Resource ReSerVation Protocol (RSVP) is a signaling protocol that can undertake this role. This book describes both the Integrated Services architecture and RSVP in detail.

[4] Odlyzko, A., "The Economics of the Internet: Utility, Utilization, Pricing, and Quality of Service," 1998. Available at:
`www.research.att.com/~amo`

A paper arguing the point of view that overprovisioning data networks is a viable and economically sustainable response to the demands for service quality within data networks, and that such a response is technically and economically superior to implementing QoS responses within the network.

[5] Braden, R., Clark, D., and Shenker, S., "Integrated Services in the Internet Architecture: An Overview," RFC 1633, June 1994.

This RFC describes the components of the Integrated Services architecture, a proposed extension to the Internet architecture, and protocols to support real-time traffic flows through service-quality commitments.

[6] Blake, S., Black, D., Carlson, M., Davies, E., Wang, Z., and Weiss, W., "An Architecture for Differentiated Services," RFC 2475, Proposed Standard, December 1998.

The architecture description for the Differentiated Services enhancements to the Internet Protocol. This architecture achieves scalability by aggregating traffic classification state, which is conveyed by means of IP-layer packet marking using the Differentiated Services (DS) field. Packets are classified and marked to receive a particular per-hop forwarding behavior on nodes along their path. Sophisticated classification, marking, policing, and shaping operations need to be implemented only at network boundaries or hosts. Network resources are allocated to traffic streams by service-provisioning policies that govern how traffic is marked and conditioned upon entry to a differentiated services-capable network, and how that traffic is forwarded within that network.

[7] Gray, T., "Enterprise QoS Survival Guide: 1999 Edition," 1999. Available at:
`http://staff.washington.edu/gray/papers/eqos22.html`

A detailed view of an approach to supporting QoS in an enterprise environment. The paper is an excellent example of the procedural steps involved in network engineering, detailing the intended environment, the available tools and the desired outcomes, and then examining the viability of a number of QoS solutions.

[8] Huston, G., "Next Steps for the IP QoS Architecture." Available at:
`www.ietf.org/internet-drafts/draft-iab-qos-00.txt`

While there has been significant progress in the definition of IP QoS architecture, there are a number of aspects of QoS that appear to need further elaboration as they relate to translating a set of tools into a coherent platform for end-to-end service delivery. This document highlights the outstanding issues relating to the deployment and use of QoS mechanisms within the Internet, noting those areas where further standards work may be required. This draft is a work item of the Internet Architecture Board Working Group of the IETF.

GEOFF HUSTON holds a B.Sc. and a M.Sc. from the Australian National University. He has been closely involved with the development of the Internet for the past decade, particularly within Australia, where he was responsible for the initial build of the Internet within the Australian academic and research sector. Huston is currently the Chief Technologist in the Internet area for Telstra. He is also an active member of the IETF, and is the chair of the Internet Society Board of Trustees. He is author of *The ISP Survival Guide,* ISBN 0-471-31499-4, *Internet Performance Survival Guide: QoS Strategies for Multiservice Networks,* ISBN 0471-378089 and coauthor of *Quality of Service: Delivering QoS on the Internet and in Corporate Networks,* ISBN 0-471-24358-2, a collaboration with Paul Ferguson. All three books are published by John Wiley & Sons. E-mail: `gih@telstra.net`