Pricing Framework for a Differential Services Internet

MICHAEL S. BORELLA, VANDANA UPADHYAY, IKHLAQ SIDHU
3Com Corp., 1800 W. Central Rd., Mount Prospect IL, 60056, USA
{mike_borella,vandana_upadhyay,ikhlaq_sidhu}@3com.com

Abstract. A number of recent proposals and proposed standards have addressed adding differential services to the Internet. Although their details and tentative implementations differ, most are recommending what essentially amounts to multiple levels of best-effort service. In this paper, we survey recent differential services and pricing proposals and introduce a pricing framework for a differentiated-services network that focuses on simplicity, flexibility, and ease of implementation. In particular, our model can be used in the current heterogeneous Internet without major restructuring. We focus on flat-rate, per-time and usage-based pricing, where users can change their service level on an ad-hoc basis. We also show that a sender-pays model with back-charging is simple to implement and can be effective even in complicated transactions, such as multicasting. Our approach is different from many previous proposals and consciously integrates differential services and pricing with implementation as the immediate goal. Finally, we outline future areas of research including ISP support for pricing, the dynamics of service quality and pricing in a differentiated Internet, and the latest trends and directions of QoS deployment and standardization..

1 INTRODUCTION

Currently the Internet, and IP networks in general, supports a single level of best-effort service. All data packets are queued and forwarded with the same priority. There is no guarantee that any given packet will actually reach its destination, much less arrive in a timely fashion. This flat service model has worked surprisingly well, as the Internet has grown at a near-exponential rate for almost 30 years. Most alternatives to IP, such as Frame Relay and ATM, provide facilities for defining quality-of-service (QoS) provisions; i.e., giving some traffic flows a higher priority, a lower packet loss rate, and/or guaranteed capacity. However, these protocols have not enjoyed IP's explosive growth.

The emergence of delay-sensitive[1] applications, such as Internet telephony, video-conferencing, and interactive gaming, are beginning to threaten IP's ubiquity. The inability of IP to reserve capacity and bound delay and loss often causes these applications to perform poorly. There is a large and growing user population that is frustrated with Internet latencies [1], and may be willing to abandon IP networks in favor of networks that can provide a higher quality of service.

RSVP [2] has been proposed by the IETF's integrated services working group as a method for allocating capacity at each router in a flow's path so that, in conjunction with admission control and a bounded-delay queuing discipline, such as Weighted Fair Queuing or one or its variants [3], end-to-end QoS guarantees can be made. However, RSVP is very resource intensive (in terms of signaling and control, as well as requiring per-flow scheduling) and has not enjoyed strong support throughout the networking community. In an effort to extend IP so that it can gracefully support multiple traffic types, each with different QoS requirements, the IETF is currently developing an architecture for providing differential services (DS) in the Internet. This architecture defines a number of per-hop behaviors (PHBs) that, when indicated in an IP packet header, define how a router should treat the packet. There is no connection setup, management, or teardown. No separate control stream is required and there is no additional overhead for DS, except at the edges of DS networks. The IETF does not explicitly recommend or define end-to-end services – instead it hopes to provide a rich enough set of PHBs so that arbitrary end-to-end services can be built on top of these primitives. The PHBs do not specify or recommend packet-scheduling or queue-management disciplines. Thus, it is up to the market (ISPs and router manufacturers) to create and sell differential services to the Internet community. It is possible that some providers will feature multiple levels of best-effort service with each service class mapped directly to a strict priority queuing scheme, while other will offer similar

[1] Most delay-sensitive applications are also sensitive to some degree to the second moment of delay (jitter) as well.

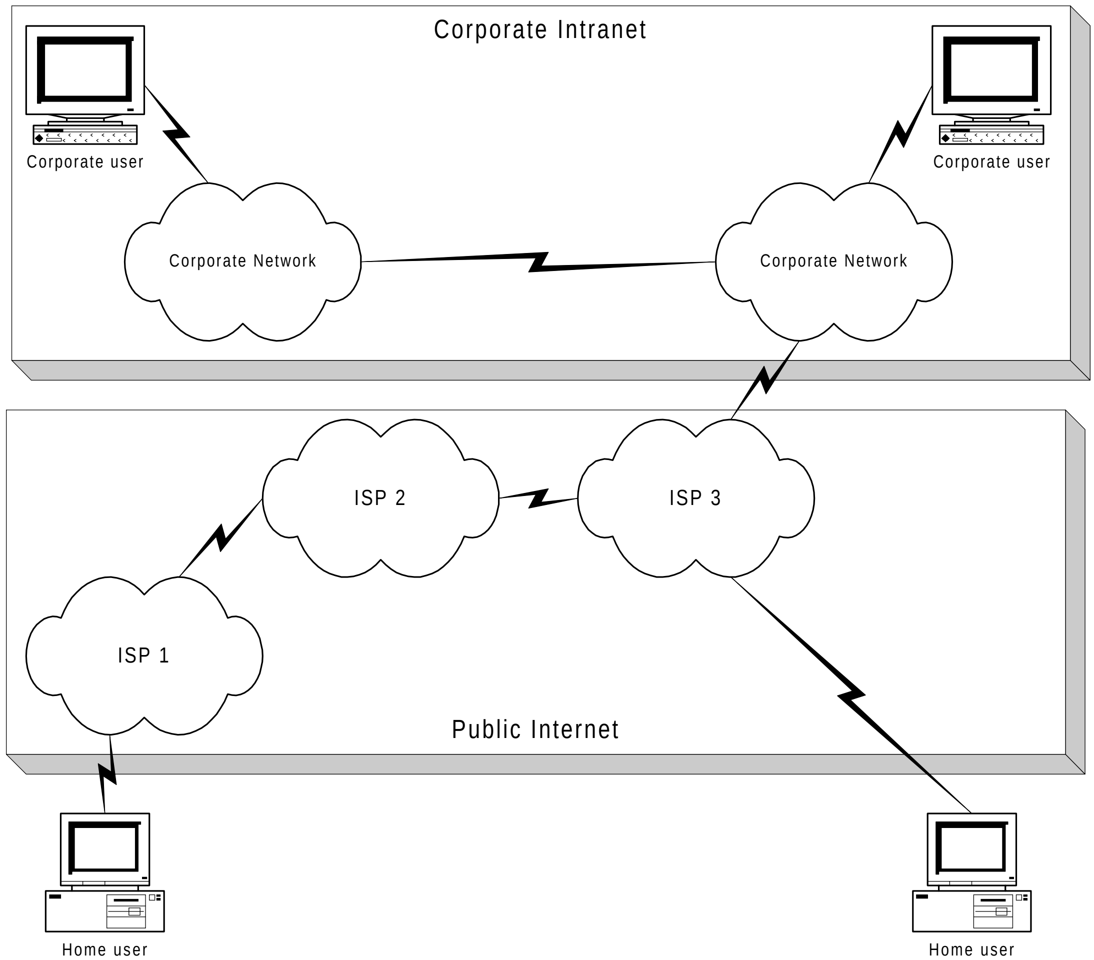

Figure 1: An example of the current Internet architecture.

services with each class mapped to a round-robin or weighted queuing scheme.

Adding support for traffic classes to the Internet is not a minor undertaking. In particular, an architecture that allows users to request higher priority for their traffic must be accompanied by an enforceable policing and pricing scheme so that users will behave in a socially acceptable fashion. Without network support for both policing and pricing user traffic, users have no reason not to always request the best possible service.

1.1 CURRENT APPLICATIONS

Traditionally, the Internet has provided only data delivery services, such as email (via SMTP), file-transfer (via FTP), remote access (via telnet or rlogin), and, more recently, World-Wide Web (WWW) access (via HTTP). Of these applications, email and file-transfer require a reasonable throughput and are loss-sensitive, but are not delay sensitive. Due to their interactive nature, remote access and web protocols are both delay- and loss-sensitive. To some extent, web protocols also require a reasonable throughput, given that many web pages contain large graphics files. All four types of services use TCP for reliable transport. TCP guarantees loss-less delivery, at the expense of additional delays caused by timeouts and retransmissions. TCP is also a conservative protocol, in the sense that it adapts its transmission rate to perceived network conditions (i.e., it will lower its transmission rate when it detects network congestion or observes events which indicate that network congestion is highly probable). Thus, TCP provides no bounds on minimum throughput or maximum delay.

Emerging applications, such as real-time, interactive voice, video, and gaming, are throughput- and delay-sensitive, but are usually not disrupted by occasional losses. Since these applications are not well suited for TCP, they use UDP instead. UDP is connectionless and provides no guarantees with respect to throughput, delay, or loss, but can transmit at a constant rate regardless of network congestion. UDP does not retransmit lost packets. The users of real-time protocols are the most vocal when QoS is poor. Roundtrip delay greater than about 250-300 ms can cause long conversational pauses that render an Internet telephony call less interactive. Delays on this order on magnitude would be common for highly-compressed voice sessions over the Internet [5].

1.2 Current Architecture

DS will not be something that is suddenly “turned on.” It must evolve within and slowly change the existing framework of the Internet. Figure 1 shows an example of the current Internet architecture[2]. Home users connect to regional or national ISPs through low-capacity dialup lines (i.e., analog modems, ISDN[3]). ISPs connect to one another through public Network Access Points (NAPs) or through private peering agreements. ISPs typically use very high capacity links, such as T3, OC-3 or OC-12. Corporations[4] connect to the Internet via their own network, which usually has only one or a small number of links to an ISP. Corporations with multiple physical sites often lease dedicated T1 or T3 lines between their locations, forming a corporate intranet. Corporations also may use a virtual private network (VPN) over the Internet to link two or more physically diverse sites.

Within the Internet model shown in Figure 1, there will exist three entities that play a role in pricing schemes: users, corporations, and ISPs. Contracts exist between users and ISPs, corporations and ISPs, and ISPs and ISPs. Currently, these contracts usually take the form of flat-rate monthly charges. For example, the typical home user pays between $15 and $30 per month for unlimited access. ISPs lease their backbone (T1, T3, on up) links from telephone companies for a fixed charge (usually on the order of several hundreds or thousands of dollars per month). ISPs must also pay a monthly rate to have access to NAPs. Private peering arrangements between ISPs may incur no charge to either party (both agree to carry the others’ traffic) or may include a one-way charge if the traffic flow is expected to be asymmetric. Usage based charging has been used in limited circumstances [6].

1.3 The Case for Differential Services

In recent years, it has become clear that the current Internet architecture cannot easily support applications with stringent delay and capacity requirements. The fact that Internet telephony and video-conferencing have not become more popular attests to this fact. Many companies are currently exploring the possibility of providing such services over a dedicated IP, Frame Relay, or ATM network rather than the Internet. Providing a real-time service, be it voice, video, or gaming, over a network with a single administrative entity allows for capacity planning and upgrading. Thus, users can be provided a reasonable quality of service through network over-provisioning rather than integrated or differential services. However, if each type of application is to have its own dedicated network, understanding the capabilities of, subscribing to, and managing these separate networks becomes a headache for corporate and ISP information technology (IT) personnel as well as home users. We would also lose any potential multiplexing gains, as one network may be congested while others have ample spare capacity.

While it may be possible to over-provision a single backbone, there are a number of drawbacks to using provisioning as the only method of end-to-end QoS:

- A single over-provisioned IP backbone can only provide QoS on itself. Typical Internet transactions require that packets traverse two or more independently administered IP backbones. Even if all of these backbones are over-provisioned, there is no guarantee that they will have the peering capacity to handle the traffic between them. Internet measurements have shown that a large amount of delay and packet loss occurs at public NAPs that interconnect several networks.
- In order to reduce congestion and maintain low end-to-end delay and negligible packet loss, a network may have to perform admission control, by dropping some packets at its ingress routers. Since this dropping occurs at peering points and/or NAPs, it is easy for a network’s administration to blame other networks or the NAP’s administration for the loss. This can result in a significant amount of loss at a peering point, but all of the connected networks reporting negligible loss. The user’s QoS will be deleteriously affected, but all of the networks are fulfilling their QoS contracts.
- Packets from all types of applications are given the same QoS. If an over-provisioned network link or router fails, or if capacity planning is not performed correctly, congestion may occur. Thus, packets with rigid delivery times may be queued behind packets with elastic delivery times, while packets that are highly loss-sensitive may be dropped before packets that are less loss-sensitive.
- Last mile access has been traditionally low bitrate. No amount of over-provisioning will allow a reasonable quality video conference to occur over a dial-up modem link. Residential broadband services, such as cable modems, DSL, and wireless-to-the-home seek to alleviate the last mile bottleneck.

[2] The ISPs are diagrammed are diagrammed as “IP clouds.” In general, the terms “IP cloud,” “administrative system” and “administrative domain” all refer to a set of interconnected IP routers and hosts that are all under the control of a single administrative entity, such as a service provider, corporation or university.

[3] As the availability of cable modems and xDSL grows, we expect more home users to connect to the Internet using these megabit-per-second technologies.

[4] Here, and in the following text, the term “corporation” could be replaced by “university” or “non-profit organization.”

Support for differential services is growing on all fronts. Microsoft plans to provide a quality of service API in Windows 2000™. The IETF is currently drafting a proposal for allowing RSVP-like signaling for bandwidth reservation on Ethernet and other networks in the IEEE 802 family [7]. The IEEE has also extended Ethernet to support eight levels of priority with the 802.1p proposal. Router manufacturers are very active in the IETF differential services forums, and have proposed several of the early DS schemes [8,9]. With this sort of strong commercial and standards support, differential services are likely to be deployed on a wide-scale basis within the next few years.

From an economic point of view, differential services will allow more efficient pricing schemes. ISPs will have the ability to charge customers and other ISPs based on the amount of each class of traffic carried. ISPs can charge higher rates to those who are willing to pay for a higher quality of service. Users of delay- and bandwidth-sensitive applications will be able to purchase the capacity that they need. Market dynamics are likely to force ISPs to be competitive so that the average home user is not "priced off" the Internet.

1.4 OUR APPROACH

This paper proposes a framework for pricing in a DS Internet. Our approach advocates the following general principles:

- No end-to-end guarantees are made by the network.
- ISP edge devices maintain per-user profiles that determine pricing and service parameters.
- Edge devices mark packets with a DS byte based on the sender's profile.
- Core routers[5] implement per-hop behavior (PHB) based on a packet's DS byte and local policy.
- Pricing is either flat-rate, per-time or usage-based.
- The user may upgrade or downgrade his service level on demand.
- A successful pricing scheme must be easy to implement by service providers and simple enough for the average user to understand.
- Pricing must be able to be added to the Internet on a network-by-network basis (i.e., a migration path must be supported).

Our goal in this paper is not to propose an "optimal" pricing model, nor are we trying to claim that our technique is the "best." In fact, we do not believe that optimal pricing (regardless of the parameter(s) being optimized) can be implemented in real-time on heterogeneous networks (see Section 3 for arguments to this effect). Instead, this paper provides a modest proposal for pricing multiple-priority traffic that can actually be implemented by ISPs. Our approach is different from many previous proposals and consciously integrates differential services and pricing with implementation as the immediate goal. This paper also serves to introduce the reader to current work in the fields of differential services and Internet pricing, and to describe the current dial-up Internet access architecture that will support pricing schemes.

This paper is organized as follows. Section 2 provides an overview of differential services and pricing schemes that have been proposed. Section 3 explains why measurement, characterization and usage-based pricing is difficult to implement in the current Internet architecture. Section 4 introduces our pricing framework for differential services. Section 5 discusses some of the peripheral issues regarding the implementation of our scheme and suggests directions for future work.

2 BACKGROUND

In this section we provide an overview of current research in differential services and network pricing. In particular, we focus on current DS proposals as well as pricing proposals. Note that in all cases, the DS proposals do not explicitly include a concrete pricing model, and in most cases the pricing proposals do not consider multiple levels of service.

2.1 IP TOS (DS) BYTE

Internet services can be differentiated by using the type-of-service (TOS) byte in the IP header (also known as the DS byte). Its format is defined in [10] as follows:

Precedence (3)	Type of Service (4)	MBZ (1)

The 3-bit precedence field uses the values 000-111 to indicate the importance of the packet. Higher values are more important, and should be have priority over lower precedence packets. The 4-bit type of service field has five defined values:

TOS value	*Interpretation*
1000	Minimize delay
0100	Maximize throughput
0010	Maximize reliability
0001	Minimize cost
0000	Normal

[5] *Core routers* are large backbone routers, generally operating at hundreds of Mbps or more. They are designed to route packets as quickly as possible. On the other hand, *edge routers* are usually only a few hops away from the user and generally do not operate at such high data rates, but instead provide some form of policy-based security, address allocation/translation, or proxy services for the user.

The 1-bit MBZ (must be zero) field is unused. The interpretation of these values and their mapping to particular router scheduling and priority queue disciplines is not defined. A drawback to this classic TOS architecture is that only one parameter from the set of delay, throughput, reliability and cost can be affected per packet. Thus, for example, a user cannot request low delay and high throughput simultaneously. In practice, most implementations ignore the TOS byte. Current DS proposals call this byte the "DS byte" and have re-defined its fields.

2.2 DIFFERENTIAL SERVICES PROPOSALS

Internet service can be differentiated in a number of ways. All implementations require the marking of the DS byte of IP packets to indicate a particular class of service. How packets of a particular class are treated by the network is a matter of policy. This section provides an introduction to the schemes that have been and are being considered by the IETF differential services working group. The fundamental architecture of all DS networks is that of traffic conditioners on the edge of the DS portion of the network, providing classification, policing and marking of packets. Once a packet is appropriately marked, it will be switched by the DS portion of the network according to its marking.

2.2.1 Early Proposals

A one-bit priority scheme was introduced in [11]. Each user has a profile that describes the characteristics of the traffic (such as mean rate, peak rate, burstiness, etc.) that the user may transmit. User traffic that fits the user's profile is marked as "in" by an ingress (edge) router. Traffic that does not fit the profile is tagged as "out." A core router will preferentially drop "out" packets when the router becomes congested. This mechanism, as it is proposed, is sender driven, but the authors discuss the possibility of a comparable receiver-driven scheme. Essentially, [11] allows users to *attempt* to utilize as much capacity as they desire. If the network is uncongested it will transport the traffic. If the network is congested, it will drop the portion of the traffic that exceeds the user's profile according to a modified Random Early Detection (RED) [12] algorithm. Overall, this proposal is similar in spirit to the discard eligibility bit in Frame Relay cells, which is set when users generate more traffic than their committed information rate.

In [8], interpretations of the DS byte values are proposed. The existing TOS and precedence bit-fields (see Section 2.1) are renamed to delay sensitivity and drop preference, respectively. The values of the 4-bit delay sensitivity field are mapped to Class-Based Queuing (CBQ) queues [13] with depths inversely proportional to the degree of delay sensitivity. The values of the drop preference field are used with a RED or modified RED scheme to preferentially drop "less important" packets.

2.2.2 Latest Proposals

Recently, the IETF DS working group has officially redefined the DS as follows [14]:

DSCP(6)	CU(2)

The 6-bit DSCP (Differential Services Code Point) field takes on a particular value that is interpreted to be a PHB. Different networks may be configured to use different code points for a given PHB. The 2-bit CU (currently unused) field must be ignored by DS routers.

2.2.2.1 Assured Forwarding

The assured forwarding (AF) PHB group [15] currently defines 12 code points: four forwarding classes, each with three drop-precedence classes. Under normal operating conditions, a DS router should forward a packet with forwarding class i before a packet with forwarding class j, if $i<j$. Likewise, in the case of congestion, a router should drop a packet with drop precedence m before a packet with drop precedence n, if $m<n$. Routers may queue packets in four separate priority queues and serve higher priority queues strictly before lower priority queues, or may implement a weighted queuing scheme, such as weighted round robin. The packet dropping mechanism may either be drop-tail or a variation of RED.

Using the assured forwarding code points, network-wide or end-to-end services can be built. For example, an ISP can sell two or more classes of best-effort service, in which some classes are better than others in a statistical sense. In other words, service A may exhibit a lower mean delay and lower mean drop rate than service B. This will allow ISPs to charge more for service A without having to provide hard QoS guarantees.

2.2.2.2 Expedited Forwarding

The expedited forwarding (EF) PHB [16] specifies a class of packets that a DS router must forward with a particular rate, independent of the forwarding rate of all other PHBs implemented at that router. The implementation can be as simple as a priority queue that is always serviced before all other queues.

The EF PHB can be used to provide a virtual leased line service. A traffic source negotiates with one or more DS networks for the dedicated bit rate that it requires. If the negotiation is successful, the source is given permission to transmit up to the bit rate. If inter-network service is offered, the guaranteed bit rate must be allocated through all of the edge routers between the source and destination. A lightweight RSVP-like signaling mechanism for doing so called "bandwidth brokers" is discussed in [9].

In order for EF to guarantee that the aggregate rate of EF packets forwarded from a given router is greater than the aggregate arrival rate at that router, a number of traffic engineering mechanisms must be in place. In particular, (1) the DS network must be over-provisioned with respect to the EF traffic that it admits, (2) admission control must be performed on the edges of the DS network, and (3) edge routers must shape (rate limit) all EF sources such that their maximum negotiated bit rate is never exceeded. Experimental results have shown EF to provide a low-loss, low-delay, low-jitter end-to-end service. In many ways, the virtual-leased line service enabled by EF is an amalgam of three different QoS mechanisms: differential services, integrated services, and over-provisioning.

2.2.3 Perspective

Since DS is not yet a mature technology, few other PHBs have been proposed. Furthermore, as of the time of this writing, DS has not been widely implemented and the AF and EF PHBs have not yet been approved as standards-track RFCs. However, the services that can be provided through the use of the AF PHBs are simple for users to understand and likely to be deployed rapidly. Due to this fact, we focus on AF-enabled services for the remainder of this paper.

2.3 PRICING PROPOSALS

In computer networks, a user is charged for network access based on one or more of the following three disciplines:

- Flat-rate: The user receives unlimited network access, regardless of the amount of time connected to the network and the amount of traffic sent or received.
- Per-time: The user pays based on the amount of time connected to the network, regardless of usage during this period of time. The charge may fluctuate based on time of day or other factors. This charging model is close to that of long-distance telephony.
- Usage-based: The user pays a marginal cost (which may vary) for each packet, cell, or byte transmitted and/or received.

Combinations of the above disciplines are also possible, and have been used in practice. For example, an ISP may choose to offer flat rate service up to a predetermined number of hours per month, after which the user is charged per hour of connect time.

From an economist's point of view, a pricing scheme is efficient if no user is receiving a quality of service that another user values more but has been denied [17]. Several notable pricing proposals have been made for both IP and ATM networks. Although the ATM schemes can not always be directly applied to IP networks, they often contain useful concepts that may be applicable.

One of the first offerings in the network pricing literature was [18]. In this paper, the authors formulated the user satisfaction of priced multiple-priority services as an equilibrium point. Through simulation of this abstract relationship they found that a two-level priority/pricing scheme produces a greater overall user satisfaction level than a single-level scheme. This is due to the ability of users to be able to purchase a higher priority when they need it. They also show that as congestion grows, the price that users are willing to pay for higher priority services also grows. While these results are purely theoretical, they do confirm our intuition about pricing and they provide a basic analytical and simulation framework for studying pricing and DS issues.

In [19], a "closed-loop" pricing scheme is developed in which a per-packet price is determined by the buffer length at a local gateway – the longer the buffer, the higher the price. The price is fixed for a predetermined period, after which the buffer is sampled again and a new price is assigned, if necessary.

An alternative scheme, called "smart market pricing," was introduced in [20]. In this proposal, users bid for network access on a per-packet basis by telling their local gateway (in each packet's header) how much they are willing to pay for the packet to get on to the network. The gateway admits the packets with the highest bids, but all users whose packets have been accepted are only charged the maximum bid amount of all packets that were not accepted by the gateway.

Both of these latter schemes propose congestion-based per-packet pricing. Their goal is to charge the user relative to the amount of congestion in the network. When the network is uncongested, the cost of transporting a new packet is minimal. When the network is congested, the cost of transporting a new packet grows proportional to the degree of congestion. This model implicitly assumes that there exist mechanisms for reasonably accurate measurement of network congestion, and the availability of tools with which users can determine what they are willing to pay for access. The DS model of multiple priorities is not explicitly addressed by these schemes, but it is not difficult to interpolate how they can be extended to support multiple traffic classes.

3 THE CASE AGAINST MEASUREMENT, CHARACTERIZATION, AND USAGE-BASED PRICING

Many pricing schemes implicitly or explicitly rely on measurement or characterization, and charge based on usage. For example, "closed-loop" pricing requires measurement, and some other schemes (see [21] for examples) require that the user present the network with a characterization of their traffic, to which the network responds with a price. In this section, we explain why these schemes are difficult to implement efficiently in real networks. Here, we use the networking provider's definition of efficiency; that is, we measure the efficiency of an implementation in terms of simplicity, speed, computational and storage requirements, and demands made on the user. In other words, an efficient pricing implementation will not require a large amount of overhead (in terms of configuration or staff re-training), cost, performance, or customers' perceived usability, as opposed to the current flat-rate model.

3.1 MEASUREMENT

Although there have been some remarkable advances in the state of the art of Internet measurement recently (see [22] and references within), in general Internet behavior is very difficult to capture analytically. Pricing models that recommend a varying per-packet price based on congestion will run into a number of implementation and fairness issues. In particular, the meaning of "congestion" is ill-defined in this context. From the user's perspective, "congestion" is indicative of some amount of delay, loss, or bandwidth scarcity in the network at the time at which the user is running some network application. Naturally, user perception of congestion will be subjective and application dependent – as discussed earlier an interactive gaming session requires low delays while transmitting email is not particularly delay-sensitive. Summarized below are a number of techniques that can be used to measure one aspect or another of congestion; however, none provide an adequate metric of the overall quality of service that a user will experience.

"Closed-loop" pricing (see Section 2.3) measures congestion by sampling the ingress queue length at the first-hop router. Although this method would be relatively easy to implement (say, using typical SNMP services), local queue lengths only imply local, not global, conditions. A short local queue would give the users a low price per packet, but if that user were trying to access a congested WWW site across the Internet, he would receive a poor quality of service. Conversely, local congestion at a user's ISP would cause the price per packet to be high even if the rest of the hops on her packet's path were uncongested. Both of these situations would encourage inefficient user behavior: using the network when it is congested and not using the network when it is not congested.

End-to-end congestion can be measured with ping-like applications. These programs, which are distributed with most workstations, send an echo-request ICMP packet to a target host, and wait for a response from that host. A series of ping transmissions can be used to estimate round-trip delay and loss to and from a site. However, if we were to implement a pricing scheme based on pinging remote sites, a number of problems would occur. Internet delays are highly variable, often by one or more orders of magnitude [23]. Delay measurements taken at any given point in time are correlated to, but not necessarily indicative of, the delays that users will experience in the future. Furthermore, sampling delays from point to point between many administrative domains would cause a large amount of "control" traffic to permeate the Internet. In order for *n* edge networks to sample delays to all other edge networks, *n x n* samples must be made per sampling period. Given the volatility of Internet dynamics, this sampling period would have to be small (on the order of seconds or minutes), further exacerbating the issue by increasing the control traffic overhead. All of this traffic would have to be collected, processed and stored in real-time, which would slow down routers and switches. Thus, not only can a measurement scheme create additional network traffic, it also will not be able to give a reasonable indication of network conditions in the future. Deployment of an effective, scalable wide-area measurement and monitoring system is an open issue.

3.2 CHARACTERIZATION

Characterization-based pricing is the general concept of charging based on the characteristics of the traffic that a user generates. Several theoretical models of traffic source bitrate dynamics, effective bandwidths and burstiness are summarized in [21]. These models are extensions of traditional queuing theoretic and Markovian analysis, and exhibit the advantages and disadvantages of such analytical models; that is, closed-form results can sometimes be readily obtained but their relevance to actual network traffic is always in question. A pricing model for an effective-bandwidth traffic analysis is proposed in [24].

Promising measurement-based source models were reported in [25] and [26] for FTP, telnet, SMTP, NNTP and several other applications. These models were shown to be representative of real network traffic in a rough sense. In [27], LAN traffic session length for individual sources was reported to be well-modeled with a Pareto distribution, however packet dynamics were not reported. As of the time of this writing, no pricing methods based on empirical models have been proposed.

In practice, pricing based on characterization of user traffic is very difficult. In the current Internet, the user has little control over how an application generates traffic. Some applications generate traffic based on user activity and are idle otherwise. Other applications transmit at some rate regardless of user activity. Few, if any, applications allow the user to explicitly control or shape their traffic characteristics. If TCP is used, both the user and the application have no control over how traffic is generated – TCP transmissions are usually bursty and are rate-controlled by an adaptive, feedback-based mechanism that is application independent. It is well known that increasing the burstiness of a traffic source has a deleterious impact on the queuing behavior of that traffic, but the degree to which one user's bursty traffic affects the entire network when aggregation levels are high is not understood. Furthermore, there are few widely accepted metrics for characterization, and those that are widely accepted are very general. For example, in ATM networks, all traffic is pigeon-holed into one of three categories: constant bit rate (CBR), variable bit rate (VBR) and available bit rate (ABR). While it should be possible to characterize CBR traffic, the latter two categories encompass a wide range of possible traffic types and are not easily characterized.

It is possible, however, to condition a user's traffic with a leaky-bucket or token-bucket at an ingress edge router. The rate and/or depth of these shapers can be used as a basis for pricing. The downside of this approach is that not all applications generate traffic that is well-modeled by a simple traffic shaper. A method currently under study involves dynamic renegotiation of token bucket parameters for particularly bursty sources. This method would likely have to be accompanied by a change in price for each renegotiation. As discussed in the next session, users may not find services for which the pricing changes dynamically to be appealing.

3.3 Usage-Based Pricing

In usage-based pricing schemes, users are charged on a per-packet, per-cell, or per-byte basis. While these schemes have attractive theoretical properties, their successful implementation and deployment may prove prohibitively difficult. Immediately, packet loss becomes a problem. When a user's packet is dropped somewhere in the network, does the user have to pay for the packet? Loss rates of 5%-15% or more are common in the current Internet [5].

Proponents of usage-based pricing claim that this discipline will reduce congestion by making users pay for the network capacity that they use. In particular, they argue that usage-based pricing will enable delay-sensitive applications such as telephony and video conferencing [20]. However, the advantages of usage-based pricing do not come for free – there will be significant difficulties implementing such a scheme in the current Internet. In particular, multimedia users will find that they will have to pay orders of magnitude more for a video conference as opposed to textual email.

Many video coders transmit at variable rates based on the complexity and movement of the subject. This would result in users with a static background paying less than users with a busy background. Also, not all packets are created equal – some packets are more important to user perception of application quality than others. For example MPEG I-frames are more sensitive to loss than MPEG B-frames or P-frames. Furthermore, not all packets contain the same amount of information with respect to their size. Packets with compressed payloads will contain more information than many uncompressed packets of the same size in bytes.

Finally, it is not clear that usage-based pricing will be attractive to users. Network users think in terms of services, not packets. Usage-based pricing would be a difficult mental barrier for users to overcome; i.e., for two different 10-minute video conferences between the same points, users will almost always have to bear dissimilar costs. See Sections 4.5 and 4.6 for a discussion of the challenge of implementing usage-based pricing.

4 A Differential Services Pricing Framework

In this section we introduce a framework for pricing differential services. We have attempted to make this framework distinct from actual pricing implementations. In other words, the framework provides an architecture for pricing, but the pricing policies that network providers use may vary. Our goal for this framework is simplicity. From a user perspective, the pricing scheme must be easily understood so that the user feels comfortable with it. A simple scheme will also make the implementation (security, accounting, policing, and marking) easier for ISPs.

4.1 What the User Sees

- *Best-Effort Delivery*: Supporting guaranteed delivery or bounded delays in an IP network is generally considered impossible without either over-provisioning the network or setting up end-to-end virtual circuits that reserve capacity on all links and routers between the two endpoints. RSVP and the EF PHB both provide a framework for the latter, but pricing mechanisms for either have not been developed, nor have these schemes been widely deployed. Naturally, it can be argued that without end-to-end guarantees, even the users with the highest service classes will not receive sufficient quality of service. However, with appropriate traffic engineering, it is possible that the network will reach an equilibrium at which users with the highest class of service will receive a quality of service good enough for real-time interactive applications. A model of the dynamics of a two-tier best-effort network is introduced in Section 5.1.
- *Multiple Levels of Service*: The user has two or more default classes of service to choose from. Only one class is used at any one time per application or user. Pricing is based on the class used; i.e., the higher the class, the higher the price and the higher the priority that the user's traffic gets in the network. A user's default service class is stored in his profile at his ISP. Service classes may also differentiate traffic based on the type of service desired. For example, a service class may include a level of delay sensitivity as well as a level of loss sensitivity. In this case, the user or

his applications can indicate the service needs of the individual packets, and multiple default service classes, based on application type, may be specified.

- *Pricing Discipline*: Pricing can be flat-rate, per-time or usage-based. As discussed in Section 3.3, usage-based pricing is more complicated to implement and may be unattractive to many users.
- *Ad-Hoc Service-Class Modification*: A disadvantage of a static user profile is lack of flexibility. Users may want to send an important email or initiate an important video conference and temporarily pay more for a higher class of service. Thus, we introduce the concept of an ad-hoc service-class modification, in which a user's traffic is given a different priority level for some temporary duration. The charging model used for ad-hoc modifications is not necessarily the same as that of the user's default service. As an example, consider a low-priority user who requires a high-quality video conference service. Knowing that his default service class is insufficient, he negotiates an ad-hoc upgrade with his ISP. For the duration of the session, all of his outgoing packets will be marked with a higher priority level. The ISP will charge him above and beyond his regular profile with either a flat rate for her entire session, or on a per-time or usage basis. Ad-hoc service downgrades are possible as well. A user who owns a high default service level may want to downgrade his service level to save money if the network can still provide a reasonable QoS at the lower level.

4.2 What the Network Does

In this section we discuss the policies that network entities use to implement the services discussed in Section 4.1. These policies are based on the recommended DS architecture [4].

- *Traffic Conditioning at the Edges*: Recall that the DS byte in the IP header determines the per-hop behavior (PHB) of the packet throughout the rest of the network. The DS byte must be marked by an entity that knows the sender's service class, such as the sender's ISP. For performance purposes, the marking mechanism should be at the point of least multiplexing, where the data rate is lowest. This implies that marking should occur at the ISP's access device (i.e., modem pool or network access server) rather than their first-hop router or Internet gateway. Marking can occur at the user's workstation, but this would require an additional policing mechanism at the ISP to guarantee that the user did not maliciously inflate her level of service.
- *Per-Hop Behavior Implementation is Determined Locally*: Once a user packet is marked and sent out to the Internet, it will pass through some number of networks, each with independent administrative entities, before it reaches its destination. These entities may or may not support differential services. If they do, implementation of service priority is locally configured. For example, ISP A may implement strict priority queueing (in which all higher-priority packets are served before any lower-priority packets) at all routers while ISP B may implement a weighted priority queueing scheme (in which at most some number or some percent of higher-priority packets are served before any lower-priority packets). Packet dropping policies, such as drop-tail, RED or preferential drop, can be implemented orthogonally.
- *Peering Agreements*: All autonomous networks that are connected must establish peering agreements with one another. These agreements are likely to be on a bilateral basis. They may vary quite a bit in complexity. The simplest peering agreement will consist of two entities concurring to carry one another's traffic, regardless of traffic type. More complex agreements will put limits on the number and service levels of packets transmitted in each direction between the entities. If one entity transcends its profile, the other may charge based on the magnitude of the transcendence. In this case, policing must be implemented at the ingress router of both entities. If the link is of a very high bit rate (100's of Mbps or more), performance will be affected if every packet is logged; thus, statistical logging based on some sampling rate (say, every 10th or 100th packet) can be used.

4.3 Service Dynamics

Within the framework discussed in Sections 4.1 and 4.2, many possible scenarios emerge. As DS is deployed throughout the Internet, we expect that some number of networks will implement DS and others will not. Thus, a given packet may not receive its marked DS priority end-to-end. However, a higher-priority packet will always receive at least the service level of a lower-priority packet.

A difficulty inherent in this framework is the possibility that packets may be re-marked by one or more networks on their way to their destination(s). The motivations for this re-marked are various. If ISP A knows that its peer, ISP B, is not policing incoming traffic, ISP A could re-mark all packets heading to ISP B with the highest service level. While this action would be socially irresponsible (it would cause all packets that pass through ISP B to suffer extra congestion and delays), it could make ISP A compare favorably to its competitors in terms of throughput and delay.

Re-marking (or dropping) could also occur in the case of a peering agreement when ISP A exceeds its profile. Instead of ISP B charging ISP A based on the extra traffic received, ISP B could downgrade the service class (or drop) the packets from ISP A that exceed the parameters of the peering agreement.

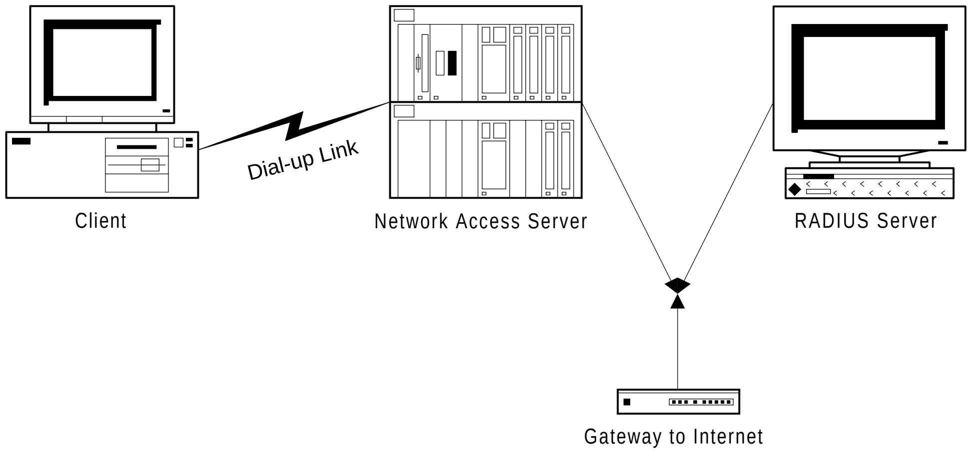

Figure 2: Typical ISP dial-up and authentication architecture.

In both of these cases, at best the user will not receive the service class that they paid for, and at worst these actions will cause further congestion. In many ways, the decision of whether or not to re-mark a packet is more administrative than it is technical. For initial deployment, ISPs may sign bilateral contracts stating that neither will re-mark the DS byte of any packet, unless the re-marking is required to avoid congestion or perform the appropriate PHB.

4.4 Support for Back-Charging

In the charging mechanism discussed so far, the sender is charged based on the priority of the packets that he sends. There is no explicit support for charging a user for the packets that he receives. Implementing such support is a fundamental issue. Ideally, a pricing framework will include a number of primitive mechanisms that can be combined to enable the complex business relationships that need to be supported by the network. As an alternative to implementing both sender- and receiver-based charging and QoS, we show how sender-based charging and QoS with off-line back-charging can support the economics of current and future electronic business relationships.

There has been a debate in recent years over the sender versus receiver charging model. For example, if a dialup user accesses a web page, who should pay for the transaction? The user presumably benefits from receiving the data. However, the user often does not know the value or the volume of the data until it is downloaded. Furthermore, many web sites contain advertisements that slow down the user's access to the actual information – users may not be willing to explicitly pay for these ads. On the other hand, if the sender pays for all transmissions, then web page hosts must pay to distribute information that their clients might otherwise purchase. Sender-based charging is relatively easy to implement in comparison to receiver-based charging or some combination of both [28].

Using a sender-pays paradigm with off-line back-charging allows a simple, yet fair, pricing mechanism. Back-charging already exists in the Internet: providers of valuable information require users to pay a monthly or per-usage fee to access that information. Many real-time stock quotes, on-line news service subscriptions, and technical support sites require a user to purchase a monthly service agreement. Back-charging allows us to implement complex combinations of sender and receiver based charging, while only implementing sender-based charging in the network. This will provide a low-overhead alternative to proposed sender/receiver charging schemes and is simple to extend to multicast sessions.

The sender-pays model applies to relative QoS. A commercial content provider may purchase a profile from its ISP with a high class of service so that its information will be delivered to clients faster than its competitors'. Users browsing the web will purchase a profile that suits their own delay-sensitivity and budget. Audio and video conference users are likely to require a higher level of service than other users. This need can be facilitated dynamically with an ad-hoc service upgrade (see Section 4.1).

4.5 Dial-up Support for Pricing

A typical ISP dial-up and authentication architecture is shown in Figure 2. The user at the client host establishes a dial-up connection with the network access server (NAS). The NAS is connected to a RADIUS authentication server [29] and an Internet gateway by a LAN. The NAS is a chassis that contains a number of modem server cards, a network management card and a routing card. A RADIUS server contains a database of user profiles that can be the basis of authentication and charging mechanisms. A typical profile contains a userid/password pair, default serial-line protocol, and maximum transfer unit (MTU). RADIUS is an extensible, upgradeable protocol.

A dial-up connection is established as follows. The client dials into the ISP and initiates a (Point-to-Point Protocol) PPP connection with the NAS. During the PPP setup and negotiation, the client transmits a userid and password in response to the NAS login request. The NAS passes this information to the RADIUS server, which authenticates the user in the local user database. At this time, the RADIUS server may pass administrative information back to the client and/or the NAS. The RADIUS server is also notified by the NAS when the user's connection is dropped. RADIUS servers log and timestamp all such activity.

RADIUS servers are a natural place to store a user's default service class(es) and other pricing information. However, since all packets that the client generates do not pass through the RADIUS server, the service class information must be given to a device that can mark packets with the appropriate DS byte. This device can either be the client itself, the NAS, or the ISPs egress gateway.

Currently, RADIUS servers only log the beginnings and the ends of sessions. Thus, flat-rate and per-minute pricing are easily supported by the RADIUS architecture, but per-packet pricing and ad-hoc upgrades are not. Some RADIUS servers optionally allow the NAS to send periodic updates. It is possible to extend RADIUS to support ad-hoc service-level modifications as follows. When the user indicates that an ad-hoc modification is to begin, the NAS transmits an update to the RADIUS server. The update is logged (the server may also confirm or deny the upgrade to the NAS and user). From that point on, the NAS stamps the user's packets with the appropriate new level of service. When the ad-hoc modification is terminated, the NAS again notifies the RADIUS server, and the server logs the message. From post-processing of the logs, the amount of time spent at each service level can be inferred. However, usage-based pricing is not as easily supported. It may be possible for a NAS device to store the number of packets that have been stamped at each level, per user. Then RADIUS can be extended to accept periodic messages than transfer these numbers to the RADIUS server for logging. However, this solution would require a re-design of NAS software, as well as a RADIUS extension.

4.6 Campus Support for Pricing

Unlike dial-up access, corporations and universities usually have no inherent authentication mechanism beyond that of password challenges for logging onto local and remote workstations. Support for pricing can be added to these campus environments in a number of ways. In this section, we provide some examples.

Typically, a corporate or university campus consists of a number of Ethernet or Token Ring LANs interconnected with one another and/or a campus backbone (such as an FDDI ring, ATM network or Gigabit Ethernet). In most cases, there is only one gateway (router) to the Internet, which may or may not act as a firewall. The entire network is usually administered by an IT department. Some local administration may be done by individual departments.

It is likely that the corporation or university will set some default level of service (DS byte) for all outgoing packets. It will allow certain privileged users to request a higher service level, perhaps as an ad-hoc upgrade. These users will run an authentication application to request the service change from their first-hop router. Once they are authenticated, the router will mark all of their packets with an appropriately upgraded DS byte. Router management software such as SMNP can be extended to support the lookups, marking and logging. Charges generated by individual users could be billed to their departmental budgets. Similar to the case of dial-up access, logging flat-rate and per-time pricing mechanisms will require less computational time and storage than usage-based methods.

4.7 The Need for Security

In any network, the need for security cannot be underemphasized. The exploits of "hackers" and "crackers" have become the topics of mainstream media. Any pricing mechanism will be a major target for hacking because it controls some form of monetary content. A complete security analysis of this or any pricing scheme is beyond the scope of this paper – security problems are often caused by implementation errors as well as poor architectural design. However, a good pricing scheme must provide accountability in case of charging disputes between a customer and ISP. This requires a method of securely logging user session data. In Section 4.5, we explained how the use of a RADIUS server can provide logging. Given that the RADIUS server will be a potential target of attacks, ISPs must pay extra attention to securing all traffic coming from and going to the server, as well as the server itself.

The presence of security issues also argues for the use of simple pricing architectures and protocols. History has shown us that the more complicated a system is, the more likely it is to have security holes [30]. For example, the *Sendmail* program, which sends, manages and delivers email on most UNIX workstations is extremely complex (it is the sole topic of a 1000-page book! [31]) and has been the target of a large number of attacks. Smaller programs have fewer states and less code, making it easier to test for and find security problems. Complex pricing systems (such as usage-based or receiver-based pricing) will be more likely to have security problems than the simpler schemes proposed in this paper.

5 Discussion and Future Work

In this section we discuss the QoS issues that will have to be addressed in the future by those who provide differential services.

5.1 ISP QoS Dynamics

An open question that the networking community must soon face is under what conditions a multiple-priority best-effort service can provide acceptable end-to-end QoS. In order to understand the quality of service dynamics inherent in a multiple-priority best-effort scheme, we develop an illustrative model. We assume that within some geographic location there are an arbitrarily large number of users, each wanting Internet access (though not all at the same time). There also is an arbitrarily large number of identical ISPs serving these users. Each ISP offers two classes of service, class 0 and class 1 (we only consider two service classes for purposes of simplicity – this model can be generalized to multiple service classes). Class 1 traffic is always forwarded before class 0 traffic, but costs significantly more. We assume that there is a single delay- or bandwidth-constrained application, such as video-conferencing or gaming, that requires class 1 service in order to be of acceptable quality. Users will have exactly one ISP account at any time, and we assume that switching between ISPs costs the users nothing and occurs instantaneously.

The state diagram for our model is shown in Figure 3. All class 1 users can be either be satisfied or dissatisfied with their QoS (for purposes of this example, we ignore class 0 users, implicitly assuming that they are not running applications that require a particular QoS). Satisfied users will use the service at some rate. When the amount of class 1 traffic is greater than the ISP can handle, the class 1 users will become dissatisfied. This is indicated in Figure 3 by the arc labeled “Class 1 demand increases.” When users are dissatisfied, either the users or the ISP will take action. The ISP’s actions can be to either increase its capacity by purchasing more capacity, raise their prices for class 1 services, or both. The users will do one of three things:

- Cancel their current ISP subscription and subscribe to a competitor who can support their QoS needs.
- Use class 1 services less, or use them when the network is relatively less congested (i.e., overnight hours).
- Downgrade their account to class 0.

As shown in Figure 3, all of these ISP and user actions cause the ISP to become less congested, moving class 1 users back to a satisfied state. User actions occur on a fine-grained timescale (seconds, minutes or hours) while ISP actions occur on a course-grained timescale (days, weeks, months).

As the number of Internet users, as well as the amount of traffic each transmits, has grown dramatically since the network’s inception, we do not expect that the model presented above has a steady-state solution. Instead, the ISP must continually monitor user satisfaction and raise prices or upgrade when it is appropriate.

5.2 Pricing QoS in a Heterogeneous Internet

The Internet of the future is expected to be extremely heterogeneous. Different networks will be using different mechanisms for QoS. For example, a user packet may begin on a workstation tagged for a particular QoS by an application or API. It gets transmitted on an Ethernet using the IEEE 802.1p priority scheme. It is then routed to

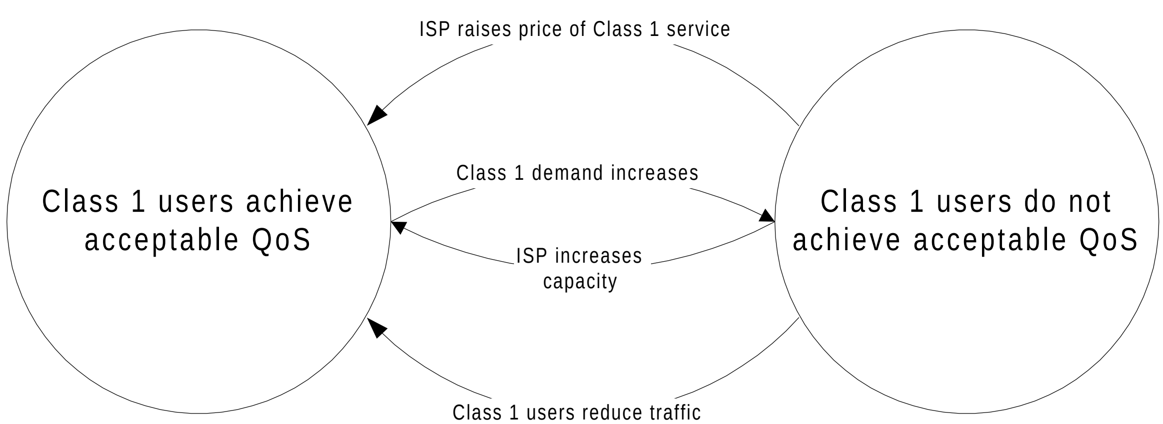

Figure 3: Flow dynamics of QoS for Class 1 users.

a DS IP backbone, which passes it off to an ATM cloud. The ATM network finally delivers the packet to its destination via a Frame Relay link.

Given the vast number of possible priority schemes that can be supported by these networks and their protocols, there needs to be agreed-upon mechanisms to facilitate translation of priorities between all possible combinations of QoS architectures. While individual mappings of service classes can be established, it will be difficult to manage as the number of priority schemes grows (i.e., mappings for N schemes).

A more desirable mapping technique will be to establish a QoS reference metric that is independent of any architecture or protocol. This metric would contain a number of queuing priority, drop precedence, and delay and jitter sensitivity classes. Then existing QoS schemes are all individually mapped to this reference metric. The challenge will be to develop a reference metric that is general enough to include all reasonable QoS implementations.

5.3 Synthesis with Integrated Services and MPLS

Currently, the future of DS is not clear. While it is certain that DS will be supported by router vendors and implemented in some backbones, DS (with the possible exception of services built using the EF PHB) may not be able to provide sufficient QoS in bandwidth constrained areas of the Internet. Last mile infrastructure and network peering points contain limited amounts of bandwidth and cannot easily be provisioned. These bottlenecks may drive a QoS architecture that includes both integrated and differential services. In particular, using RSVP to allocate bandwidth at the edges of the network and the peering points, while DS is used in the backbones has been recently proposed.

Another architecture that future DS implementations may have to combine with is Multi-protocol Label Switching (MPLS). Label switching routers (LSRs) allow one or more 32-bit tags to be inserted between end of the link-layer header and the beginning of the IP header. This tag can be used for fast routing, mapping to ATM-like virtual circuits, and any policy-based forwarding decision. Use of MPLS for DS-like services will obviate the need for IP DS support in core backbone networks. One factor, however, that will slow down the deployment of MPLS is the fact that in an MPLS-based network, *all* routers must be LSRs, while a DS network can be incrementally upgraded.

In both of these cases, pricing will be a fundamental component of the end-to-end services offered to users, and in the end, it will be up to these users to determine which QoS mechanism provides the best service. As discussed in the previous section, it is likely that different networks will support different variations of DS, intergrated services, and MPLS, and a mapping between these schemes will eventually have to emerge.

6 Conclusion

In this paper we have explored a number of possible pricing schemes for supporting differential services in the Internet. Flat-rate and per-time pricing schemes that are not based on measurement or characterization are the simplest to implement, and are likely to be the most popular with users. The main advantage of implementing DS is to give users who desire a better QoS the ability to purchase it. We have outlined a general framework for pricing DS that requires only relatively simple upgrades to current accounting hardware and software. In particular, a simple sender-based accounting scheme supports rather complex combinations of sender and receiver charging through off-line back-charging.

7 Acknowledgements

This research has benefited from discussions with Scott Jordon. The authors are also grateful to Biswanath Mukherjee, Constantinos Dovrolis, Thomas Kostas and Jerry Mahler for providing comments on an earlier draft.

8 References

[1] N. J. Lightner, I. Bose, and G. Salvendy, "What is Wrong with the World-Wide Web?: A Diagnosis of Some Problems and Prescription of Some Remedies," *Ergonomics*, Vol. 39, pp. 995-1004, 1996.

[2] R. Braden, L. Zhang, S. Berson, S. Herzog, and S. Jamin, "Resource Reservation Protocol (RSVP) – Version 1 Functional Specification," *Internet RFC 2205*, Sep. 1997.

[3] H. Zhang, "Service Disciplines For Guaranteed Performance Service in Packet-Switching Networks," *Proceedings of the IEEE*, Vol. 83, No. 10, Oct. 1995.

[4] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, W. Weiss, "An Architecture for Differentiated Services," *Internet RFC 2475*, Dec. 1998.

[5] T. J. Kostas, M. S. Borella, I. Sidhu, G. M. Schuster, J. Grabiec and J. Mahler, "Real-Time Voice over Packet-Switched Networks," *IEEE Network*, Vol. 12, No. 1, pp. 18-27, Jan.-Feb. 1998.

[6] N. Brownlee, "Internet Pricing in Practice," in *Internet Economics*, MIT Press, pp. 77-90, 1997.

[7] A. Ghanwani, J. W. Pace, V. Srinivasan, A. Smith, and M. Seaman, "A Framework for Providing Integrated Services Over Shared and Switched IEEE 802 LAN Technologies" *Internet Draft <draft-ietf-issll-is802-framework-05.txt>*, May 1998 (expired).

[8] P. Ferguson, "Simple Differential Services: IP TOS and Precedence, Delay Indication, and Drop Preference," *Inter-*

net Draft <draft-ferguson-delay-drop-00.txt>, Nov. 1997 (expired).

[9] K. Nichols, V. Jacobson, and L. Zhang, “A Two-Bit Differential Services Architecture for the Internet,” *Internet Draft <draft-nichols-diff-svc-arch-00.txt>*, Nov. 1997 (expired).

[10] P. Almquist, “Type of Service in the Internet Protocol Suite,” *Internet RFC 1349*, Jul. 1992.

[11] D. Clark and J. Wroclawski, “An Approach to Service Allocation in the Internet,” *Internet Draft <draft-clark-diff-svc-alloc-00>*, Jul. 1997 (expired).

[12] S. Floyd and V. Jacobson, “Random Early Detection Gateways for Congestion Avoidance,” *IEEE/ACM Transactions on Networking*, Vol. 1, No. 4, pp. 397-413, Aug. 1993.

[13] S. Floyd and V. Jacobson, “Link-Sharing and Resource Management Models for Packet Networks,” *IEEE/ACM Transactions on Networking*, Vol. 3, No. 4, pp. 365-386, Aug. 1995.

[14] K. Nichols, S. Blake, F. Baker, and D. Black, “Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers,” *Internet RFC 2474*, Dec. 1998.

[15] J. Heinanen, F. Baker, W. Weiss, and J. Wroclawski, “Assured Forwarding PHB Group,” *Internet Draft <draft-ietf-diffserv-af-03.txt>*, Nov. 1998 (work in progress).

[16] V. Jacobson, K. Nichols, and K. Poduri, “An Expedited Forwarding PHB,” *Internet Draft <draft-ietf-diffserv-phb-ef-01.txt>*, Nov. 1998 (work in progress).

[17] J. K. MacKie-Mason, L. Murphy and J. Murphy, “The Role of Responsive Pricing in the Internet,” in *Internet Economics*, J. Bailey and L. McKnight, eds. MIT Press, pp. 279-304, 1996.

[18] R. Cocchi, S. Shenker, D. Estrin, and L. Zhang, “Pricing in Computer Networks: Motivation, Formulation, and Example,” *IEEE/ACM Transaction on Networking*, 1993.

[19] J. Murphy and L. Murphy, “Bandwidth Allocation by Pricing in ATM Networks,” *IFIP Transactions C: Communication Systems*, Vol. C-24, 1994.

[20] J. K. MacKie-Mason and H. Varian, “Pricing the Internet,” in *Public Access to the Internet*, ed. B. Kahin and J. Keller, Prentice-Hall, 1995.

[21] S. Jordon and H. Jiang, “Connection Establishment in High-Speed Networks,” *IEEE Journal on Selected Areas in Communications*, Vol. 13, No. 7, Sep. 1995.

[22] V. Paxson, “Measurements and Analysis of End-to-End Internet Dynamics,” Ph.D. dissertation, University of California, Berkeley, Apr. 1997.

[23] M. S. Borella and G. B. Brewster, “Measurement and Analysis of Long-Range Dependent Behavior of Internet Packet Delay,” *Proceedings, IEEE Infocom ’98*, pp. 497-504, Apr. 1998.

[24] H. Jiang and S. Jordon, “The Role of Price in the Connection Establishment Process,” *European Transactions on Telecommunications*, Vol. 6, No. 4, Jul.-Aug. 1995.

[25] V. Paxson, “Empirically-Derived Analytical Models of Wide-Area TCP Connections,” *IEEE/ACM Transactions on Networking*, Vol. 2, No. 4, pp. 316-336, Aug. 1994.

[26] V. Paxson and S. Floyd, “Wide Area Traffic: The Failure of Poisson Modeling,” *IEEE/ACM Transactions on Networking*, Vol. 3, No. 3, pp. 226-244, Jun. 1995.

[27] W. Willinger, M. S. Taqqu, R. Sherman, and D. Wilson, “Self-Similarity Through High Variability: Statistical Analysis of Ethernet Traffic at the Source Level,” *IEEE/ACM Transactions on Networking*, Vol. 5, No. 1, pp. 71-86, Feb. 1997.

[28] D. D. Clark, “Combining Sender and Receiver Payments in the Internet,” *Proceedings, Telecommunications Research Policy Conference*, Oct. 1996.

[29] C. Rigney, A. Rubens, W. Simpson, and S. Willens, “Remote Authentication Dial In User Service (RADIUS),” *Internet RFC 2138*, Apr. 1997.

[30] S. Garfinkel and G. Spafford, “Practical UNIX and Internet Security,” *O’Reilly and Assoc.*, 1996.

[31] B. Costales and E. Allman, “Sendmail 2nd ed.,” *O’Reilly and Assoc.*, 1997.