

A Comparative Study of Schemes for Differentiated Services

Anindya Basu

Bell Laboratories, Lucent Technologies
basu@research.bell-labs.com

Zheng Wang

Bell Laboratories, Lucent Technologies
zhwang@dnrc.bell-labs.com

Abstract

In the past year, there have been multiple proposals for providing differentiated services over the Internet. In this paper, we do a comparative evaluation of three proposals for differentiated services – the RIO scheme, the two-bit scheme and the User-Share Differentiation (USD) scheme. We simulate the three schemes for a number of different scenarios. The results show that matching the network bottleneck bandwidth to the expected bandwidth profiles is an important issue for the RIO and the two-bit schemes. These schemes perform well when there is no such mismatch; however, if the bottleneck bandwidth available is in excess of the aggregated expected bandwidths, the bandwidth allocation is not commensurate with the user expectations. On the other hand, the USD scheme is able to do a fair allocation of bandwidth that is in keeping with user expectations, but at the cost of greater implementation complexity at the network core. We also examine the bandwidth allocation under the three schemes for short duration traffic (such as web traffic). We conclude that in order to allocate bandwidth fairly for short lived traffic, the RIO and two-bit schemes require some modifications such that they can react faster to changing network conditions. The USD scheme, however, does not depend on admission control at the network boundaries, and is therefore able to allocate bandwidth fairly for short duration traffic. We also study the performance in the presence of non-responsive sources: the RIO scheme performs better than the two-bit scheme when the non-responsive sources are CBR, however, malicious sources that flood the network are able to starve low priority traffic. The USD scheme is able to provide better traffic isolation under such circumstances. Finally we study the case where the network is under provisioned with each source sending traffic at the same high rate – in this scenario, the RIO and the USD schemes allocate the available bandwidth evenly among all the active sources, but in the two-bit scheme case, an under provisioned network may deprive some of the sources from getting any bandwidth. One implication of our findings is that some form of coarse grained signaling for aggregated flows will be necessary in order to provision a network in a more efficient fashion.

1. Introduction

The Internet is currently based on the best-effort model where all data packets are treated equally and the network tries its best to ensure reliable data delivery. The best-effort model has been successful till now because a high proportion of the traffic on the Internet is TCP-based [14]. The TCP end-to-end congestion control mechanisms force the traffic sources to back off whenever congestion is detected in the network. However, such dependence on the end systems' cooperation is becoming increasingly unrealistic. Given the current best-effort model with FIFO queuing inside the network, it is relatively easy for non-adaptive sources to gain greater shares of network bandwidth and thereby starve other, well-behaved, TCP sources. For example, a greedy source may simply continue to send at the same rate when faced with congestion while other TCP sources back off. Even today, many applications such as Netscape web browsers take advantage of the best-effort model by opening up multiple connections to web servers in an effort to grab as much bandwidth as possible. The best-effort model is also inadequate for applications such as real time audio and video that require explicit bandwidth and delay guarantees. Moreover, the best-effort model treats all packets equally once they have been injected into the network. Thus, it is difficult for Internet Service Providers (ISPs) to provide services that are commensurate with the expectations of consumers who are willing to pay more for a better class of services.

Over the last ten years, considerable effort has been made to provide Quality-of-Services (QoS) guarantees in the Internet. Much of the work has focused on the end-to-end per-session reservation approach where an application makes an end-to-end reservation before starting a session [17]. While it makes sense to set up a reservation for a long-lasting session, this approach is not suitable for short duration, transaction-oriented applications such as browsing the World Wide Web where the overheads and latency involved in setting up a reservation for each session could become disproportionately high. Furthermore, there are also concerns regarding the scalability of providing guarantees at the per-session level in the core of the Internet.

The above issues have led to a number of proposals for providing differentiated services in the Internet [1]. In contrast to the per-session reservation approach, the differentiated services approach allows service providers to offer different levels of services to a few (typically, a small number such as two or three) classes of aggregated traffic flows. For example, an ISP may offer

two levels of services – a premium service for customers who are willing to pay more and a best-effort service at a lower price. The differentiated services approach attempts to allocate bandwidth for aggregated flows without per-session signaling or maintaining per-session state in the network core, and to offer better services for users willing to pay more with long-term service contracts.

Some of these proposals, such as the RIO scheme by Clark *et al* [2] and the two-bit scheme by Nichols *et al* [11], advocate pushing traffic policing and admission control to the edge of the network and leave the core of the network as simple as possible. Incoming traffic from a user is monitored at the edge of the network against the profile that the user has registered (and paid for) with the service provider. The packets that are out of profile are either shaped or tagged and the core routers process such tagged (or untagged) packets according to predefined policies. Such an approach has the advantage of simplicity since core routers do not need to store state for each traffic flow; instead, they can process packets using different policies for each traffic class which is determined by a small number of bits in the packet header. On the other hand, since the traffic is only policed at the edge, substantial over-provisioning in the core of the network may be required to ensure that there is sufficient bandwidth for the aggregated flows. In addition, as the core of the network can only see aggregated traffic, fairness among individual traffic flows belonging to the same class may not be maintained when congestion occurs.

A different set of proposals that is exemplified by the User-Share Differentiation (USD) scheme proposed by Wang [15,16] advocate putting modestly sophisticated control mechanisms in core routers to provide proportional fair sharing (PFS) of resources and per-prefix traffic isolation. The advantage of such schemes is that they do not require a priori estimation of resources that a particular user or an application expects to use. At any point in the network, the USD scheme can ensure that the bandwidth allocated to a flow is commensurate with some predefined metric, such as the price paid by the originator of the flow to the ISP. For example, it is possible to guarantee that user A who pays twice as much as user B, gets twice as much bandwidth as user B anywhere in a network. The major disadvantage of such schemes lie in the increased complexity of the routers in the network core and the concern that such complexity may not be amenable to scaling.

In this paper, we present a comparative study of the three proposals for differentiated services, namely the RIO scheme proposed by Clark *et al*, the 2-bit scheme proposed by Jacobson *et al* and the User-Share Differentiation (USD) scheme proposed by Wang. We implement each of these schemes on a simulated network using the REAL network simulator [9] and study their performances under different network configurations. We find that the USD scheme provides a better match of performance to user expectations and is more robust with respect to short-lived flows such as web traffic, but this comes at the added cost of implementing PFS at the network core.

The rest of the paper is organized as follows: Section 2 describes the three schemes in detail. Section 3 evaluates the performance of the three schemes under different scenarios, followed by Section 4 that summarizes the results of the evaluation. We conclude the paper in Section 5 with a discussion of some of the implications of our findings.

2. Three Proposals for Differentiated Services

In this section, we describe in more detail the three differentiated services schemes, namely, the RIO scheme, the two-bit scheme and the USD scheme.

2.1 The RIO Scheme

The RIO scheme (RIO stands for Random Early Detection with In/Out bit) uses some form of packet tagging to indicate the drop priority of the packet to the core network routers. Each user (or traffic flow, depending upon the granularity) is assigned a service profile by the ISP based on the expected bandwidth utilization by the user. At the edge of the network domain managed by the ISP (i.e., at the ingress points), user traffic is monitored by a profile meter to ensure that it stays within the profile. Any packets that are out of profile are marked as “out” while those that conform to the user profile are marked as “in”. In the network core, the “in” and “out” packets are treated with different drop priorities using RED (Random Early Detection). In short, “in” packets start being dropped only when the queue size crosses a higher threshold than in the case of “out” packets and get dropped with a lower probability than “out” packets. This ensures that

in-profile traffic has less chance of getting dropped than out-of-profile traffic, and therefore gets predictable levels of service so long as it stays within profile (even when congestion occurs). The RIO scheme also makes it possible to use statistical multiplexing to utilize any excess bandwidth that may be available since it does not prevent out-of-profile packets from entering the network.

It is important to note here that the RIO scheme requires the profile meters to estimate the instantaneous sending rates for sources. To do this, the authors propose a Time Sliding Window (TSW) scheme for TCP traffic that provides an exponentially averaged estimate of the TCP sending rate over time. The TSW scheme estimates the instantaneous sending rate upon each packet arrival and decays the past history over time in order to adapt to changing network conditions. If the estimated sending rate on packet arrival exceeds a certain threshold, the packet is marked as out of profile, otherwise it is in profile. The details are described in [2].

2.2 The Two-Bit Scheme

The two-bit differentiated services architecture combines the RIO scheme with Jacobson's premium service [8] to come up with three different classes of service, namely premium, assured and best-effort. The premium service is based on explicit resource reservations and a simple priority queuing model where the premium traffic is transmitted prior to other traffic. Such a service is useful for time-critical applications that require minimum delay as well as applications that require explicit bandwidth and delay guarantees. The premium traffic is strictly policed at the edge of the network: packets that are out of the service profile are dropped or delayed until they are within the service profile. The assured service class corresponds to the "in" traffic in the RIO scheme and has a lower drop priority than the ordinary best-effort service. When congestion occurs, the chances are high that the assured packets would receive predictable levels of service if they conform to their service profile, which is typically based on average bandwidth utilization. This service is appropriate for users who are willing to pay somewhat more in order to get better (or predictable) performance in times of congestion.

In order to implement the two-bit scheme, the authors propose two distinct bit patterns in the packet headers that act as tags – the P bit for the premium service and the A bit for the assured service. Routers in the network are classified into two types, the edge routers that implement

some form of profile metering and the core routers that use the packet tags to assign priorities to packets. The core routers do not need to store any per-session state: they only need to know how to process packets with the P tag, packets with the A tag and packets with no tags. The profile meter on an edge router uses a leaky bucket algorithm to meter premium and assured packets. For premium packets, the leaky bucket consists of tokens that fill up the bucket at the peak rate that has been reserved for the premium service. When a new packet arrives, it is passed on to the forwarding engine or dropped depending on whether a token is available or not. In the case of assured packets, the leaky bucket is filled at the average rate based on the expected user profile. On the arrival of an assured packet, the A bit is reset if there are no tokens available and the packet is sent to the forwarding engine.

The forwarding engine is common to both the leaf and core routers. It implements two sets of priority queues, a high priority queue for premium packets and a low priority queue for assured and best-effort packets. The high priority queue is always serviced before the low priority queue. The low priority queue is serviced using the RIO scheme, treating the packets with the A bit set as “in” packets and the rest as “out” packets. As in the RIO case, this ensures that the assured traffic flows get predictable levels of service when they stay within profile.

2.3 The USD Scheme

The USD scheme takes an approach that is different from that of the RIO and two-bit schemes. First, instead of taking a minimalist approach with a small number of distinguishable classes, the USD scheme allows traffic isolation on a per-customer basis. This allows more sophisticated service contracts to be supported. Second, in contrast to the RIO and two-bit schemes where the traffic policing happens at the edge of the network, the USD scheme manages bandwidth allocation directly on the bottlenecks where the congestion takes place. Finally, while both RIO and two-bit schemes primarily deal with outgoing traffic traveling from customers’ networks to their ISP, the USD scheme works with traffic of both outgoing and incoming directions.

The USD scheme introduces two terms for bandwidth allocation, user and share. The *user* refers to the customer to which the bandwidth is allocated. In the USD scheme, a user can be a network, a group of networks identified by a prefix, or an individual end user. Each user is assigned a

number called *share*, based on some predefined metrics, such as the amount that a user has paid for the service. The USD scheme operates with the *proportional fair sharing (PFS)* principle, where at any point in a network, the bandwidth allocated to the traffic from or to a user is in proportion to the user's share. In other words, the exact amount of bandwidth allocated to flow from or to a user is determined by the total bandwidth available at the bottleneck and the relative weights (or shares) of active users competing for the bottleneck resources.

Proportional allocation ensures that the bandwidth allocation is always fair with respect to user shares. In the presence of different traffic dynamics, such an approach proves to be very flexible. For example, as the traffic volume fluctuates throughout the day, the PFS scheme can adapt to the changing conditions: while the actual amount of bandwidth that a user gets depends on the competing traffic, its share relative to other users always remains fixed.

In a network consisting of links that have different speeds, proportional allocation can scale with the speed of the link. If an ISP has a “thin” international link to remote ISP and a “fat” local link to a neighboring ISP, proportional allocation ensures appropriate bandwidth sharing on both the bottlenecks. Such flexibility is particularly useful for transaction-oriented data applications, where end-to-end mechanisms may not react in time to ensure that the bandwidth allocations at bottlenecks are done in accordance with user profiles. When network failure occurs, even an over-provisioned network may experience congestion, for example, during link failures. Proportional allocation ensures that there is explicit policing even under such circumstances.

The USD scheme can be implemented with a classifier capable of prefix lookup (to identify flows belonging to individual customers) and a scheduler that supports PFS. Prefix lookup has been extensively studied for route lookups and several novel algorithms have been proposed for this purpose [5]. Hardware-based route lookup implementations that can support over 1 millions entries [10] exist and the same techniques can be used in the USD for identifying which customer a packet belongs to. A large number of scheduling algorithms can support PFS. Among them, weighted fair queuing (WFQ) [12] can support PFS with the most stringent fairness guarantees. Many variations of WFQ, such as Deficit Round Robin (DRR) and Self-Clocked Fair Queuing (SCFQ), have lower implementation complexity and yet provide reasonable fairness guarantees [7,13]. For example, the PacketStar Router from Bell Labs implements a variant of WFQ with 64,000 queues at 622 Mb/s speed [10].

3. Simulating Differentiated Services

In this section, we present and discuss the simulation results for the three differentiated services proposals described in the previous section. While the ideal method of doing a comparative study would be to deploy each of the three schemes on the internet, we relied on simulations instead since deploying new mechanisms on a real network without a thorough understanding of the

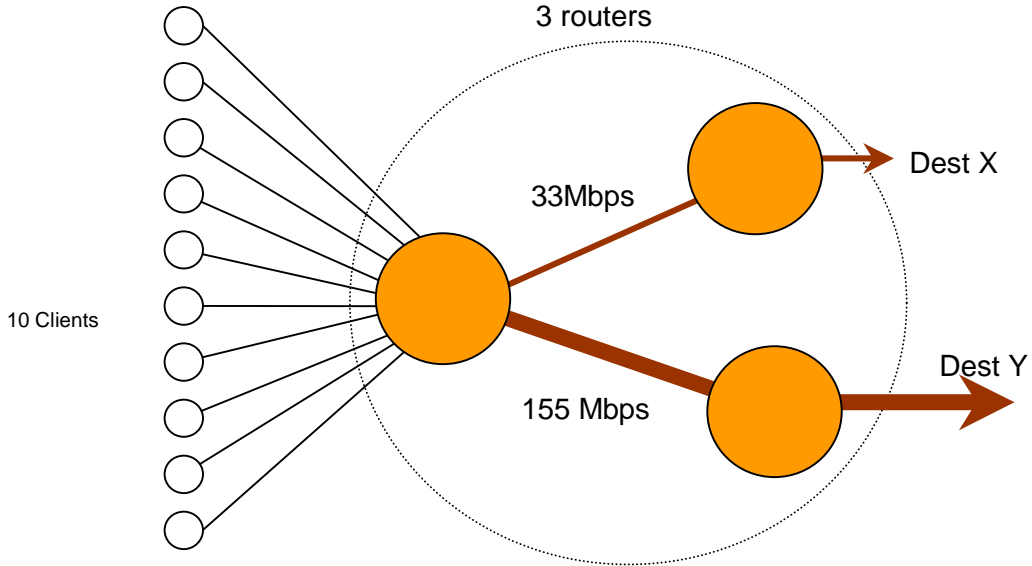


Figure 1: Simulated Network

ramifications would be inappropriate. These simulations are intended to provide us with a detailed description of how the network behaves when such schemes are deployed. Through such a process, we hope to gain a deeper understanding of the strengths and weaknesses of these schemes before they are actually deployed on a real network.

The simulations were done with the REAL network simulator [9]. For purposes of the simulation, we used a network with the configuration shown in Figure 1. In the simulation, we have 10 sources (1 through 10 counting downwards) that communicate with one of two different destinations, X and Y. The sources are connected to destination X through a bottleneck link of capacity 33Mbps and to destination Y through a bottleneck link of capacity 155Mbps. Each of the links that connects a source to the first router in the diagram has a capacity of 30Mb/s. The one-way propagation delays (the same to both X and Y) for the sources (counting from top downwards) were set at 20, 20, 40, 40, 50, 50, 70, 70, 100, and 100 milliseconds respectively.

The sources are all TCP-Reno sources (unless specified otherwise) that use a packet size of 500 bytes and each of them sends 500,000 packets to the destination. For the RIO implementation, the routers use RED with the values of 400 packets, 800 packets, and 0.02 for min_in , max_in , and P_{max_in} , and 100 packets, 200 packets and 0.5 for min_out , max_out , and P_{max_out} [2] where min_in and max_in represent the upper and lower bounds for the average queue size for in profile packets when RED kicks in and P_{max_in} is the maximum drop probability for an in profile packet when the average queue size is in the $[min_in, max_in]$ range. The min_out , max_out and P_{max_out} are the corresponding parameters for the out of profile packets. Note that the total queue sizes are used when calculating the drop probability for out of profile packets, as suggested in [2]. For the USD scheme, we used the weighted fair queuing (WFQ) discipline on the router to provide PFS. While it can be argued that WFQ was not the most appropriate scheme to use for this purpose (especially in the light of the discussion in section 2.3 which advocates use of WFQ variants with less implementation complexity and less strict guarantees), we feel that the results provide an important baseline for purposes of comparison with the other schemes.

3.1 Effect of Expected Bandwidth Profiles

We first evaluated the effect of expected bandwidth profiles and bottleneck bandwidths on actual bandwidth allocation. For this purpose, we ran three sets of simulations for each of the two bottleneck links (33Mb/s to destination X and 155Mb/s to destination Y). In each of the simulations, the odd numbered sources (1,3,5,7,9) had the lower expected bandwidth and the even numbered sources (2,4,6,8,10) had the higher expected bandwidth. In the first set of simulations, we set the expected lower and higher bandwidths to 1Mb/s and 5Mb/s, respectively, such that the aggregate (i.e., sum of) expected bandwidth of all the flows matched the 33Mb/s bottleneck bandwidth. In the second set, the expected bandwidths were chosen to be 5 and 25Mb/s, respectively to match the 155Mb/s bottleneck and in the third set, they were set to 3 and 15Mb/s, respectively, which is the average of expected bandwidths in the first two sets.

The performance of the RIO scheme is shown in Figure 1. The figure consists of six clusters with 10 bars in each cluster. Each cluster has a label of the form x Mb/s (y,z), which means that the particular cluster represents the bandwidth allocation for the 10 sources when the bottleneck bandwidth is x Mb/s and the higher and lower expected bandwidths are y and z Mb/s, respectively. The results indicate that if the aggregate expected bandwidth of all the flows passing through the bottleneck closely matches the bottleneck bandwidth, the RIO scheme is able to allocate bandwidth in accordance with the expected bandwidth profiles. This is clear from the graphs for the 33Mb/s (1,5) and 155Mb/s (5,25) cases. In the first case, the sources with a lower expected bandwidth are allocated a slightly higher bandwidth than their expected profiles at the

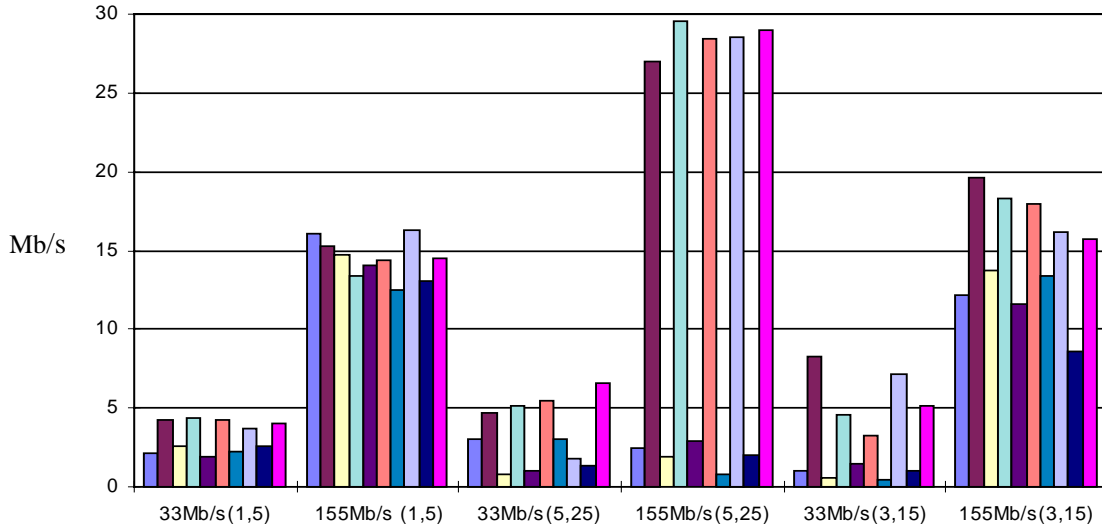


Figure 1: Bandwidth share for each of the 10 sources, for two different bottleneck bandwidths and 3 different expected profiles using the RIO scheme

expense of sources with a higher expected bandwidth and the situation is the reverse in the 155Mb/s (5,25) case. We expect that some tuning of the RED parameters is required to obtain a bandwidth allocation that is more commensurate with the expected bandwidth profiles. If the bottleneck bandwidth is in excess of the aggregated expected bandwidth, the excess bandwidth is more or less evenly allocated among all the sources. This is evident from the figures for the 155Mb/s (3,15) and the 155Mb/s (1,5) cases. In the 155Mb/s (3,15) case, the sources with the higher expected bandwidth do get more bandwidth than the others, but the relative difference is smaller than those in the 33Mb/s (1,5) and the 155Mb/s (5,25) cases where the bottleneck bandwidth matches the aggregate expected bandwidth. In the 155Mb/s (1,5) case (where the difference between the bottleneck bandwidth and the aggregated expected bandwidth is higher), the bandwidth allocation is fairly even among all the sources. This is because the excess

bandwidth is only allocated to the out of profile packets, and the RED scheme for out of profile packets ensures that this excess bandwidth is allocated fairly. Finally, if the aggregated expected bandwidth is less than the bottleneck bandwidth, the RIO scheme is able (for the most part) to allocate the existing bandwidth in a manner that is commensurate with existing user expectations. There is one exception in the 33Mb/s (5,25) case, where the bandwidth allocated to source 8 is less than that allocated to sources 1 and 7 although the latter have lower expected bandwidths. In summary, the RIO scheme performs well if the bottleneck bandwidth matches or is less than the aggregate expected bandwidth of all the flows passing through the bottleneck, but is unable to allocate bandwidth in proportion to user expectations if there is excess bandwidth at the bottleneck.

To evaluate the two-bit scheme, we set the higher bandwidth traffic as the premium traffic. For example, in the 33Mb/s (1,5) case, the 5Mb/s sources were all considered premium sources with a peak sending rate of 5Mb/s. Network bandwidth for premium traffic was reserved at connection set up time. The low bandwidth sources were subjected to the RIO scheme where all the in profile packets had their A-bit set while the out of profile packets did not. The RIO parameters in this case were the same as in the previous experiment. Figure 2 shows the bandwidth allocation

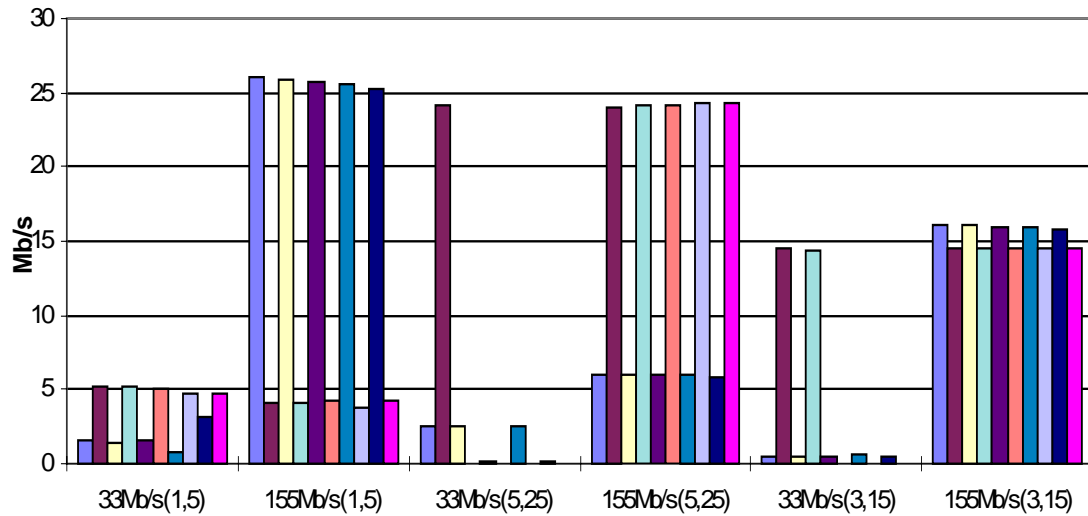


Figure 2: Bandwidth share for each of the 10 sources, for two different bottleneck bandwidths and 3 different expected profiles using the two-bit scheme.

for the two-bit scheme. We see that the bandwidth allocation matches the expected bandwidth in the 33Mb/s (1,5) case and the 155Mb/s (5,25) case. This implies that the two-bit scheme performs well when the bottleneck bandwidth matches the aggregate expected bandwidth.

However, we see that in the 155Mbit/s (1,5) and the 155Mb/s (3,15) cases, the bandwidth share allocated to the sources with the lower expected bandwidth is higher. This is because of how premium traffic is handled in the network: any out of profile premium packet is dropped by the profile meter. Consequently, all the excess bandwidth in these cases goes to the sources with lower expected bandwidth. In this respect, the bandwidth allocation done by the RIO scheme is more in keeping with user expectations. In the 33Mb/s (5,25) and 33Mb/s (3,15) cases, some of the high expected bandwidth sources do not get any bandwidth at all. This is because their connection set up calls get rejected due to unavailable bandwidth. Thus, for the two-bit scheme with premium traffic, it is important that the aggregated expected bandwidth through the bottleneck match the bottleneck bandwidth, else, some premium traffic may get denied bandwidth because the network resources have been over committed.

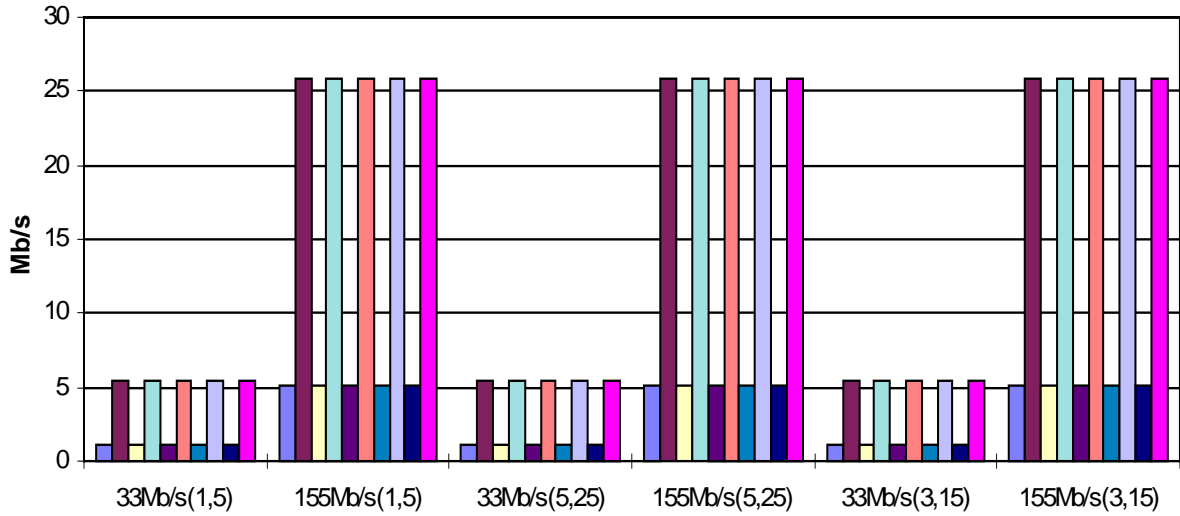


Figure 3: Bandwidth share for each of the 10 sources, for two different bottleneck bandwidths and 3 different expected profiles using the USD scheme

Finally, we simulated the USD scheme using weighted fair queuing in the network core. The bandwidth allocation is shown in Figure 3. In this case, we see that the relative bandwidth allocated to all the sources remain the same in all cases. In other words, the sources with high expected bandwidth (which presumably have paid more for such a privilege) get a higher relative share of the bandwidth than the others, independent of the bottleneck bandwidth or the absolute values of the expected bandwidth profiles.

3.2 Short Duration Traffic

A significant component of Internet traffic today is web traffic [14] that is characterized by short connection lifetimes. In order to study the effect of short connection lifetimes on the three different schemes, we ran simulations with the same settings as before except that the sources now sent only 200 packets to the destination. This corresponds to a web transfer of 100Kbytes, a moderately sized web page with embedded graphics. Figure 4 shows the performance for the 33Mb/s (1,5) (or matching bottleneck bandwidth) scenario and Figure 5 shows the performance for the 155Mb/s (1,5) (or excess bottleneck bandwidth) scenario. The numbers for the 33Mb/s (1,5) scenario show that on such short time scales, the bandwidth allocated by the RIO scheme is not commensurate with the expected bandwidth profiles. This is because the RED technique used in the routers for estimating average queue lengths (that determines when to drop a packet) and the Time Sliding Window technique used by a profile meter to estimate the sending rate for a source (that determines whether a packet is in or out of profile) depend on long term averages. In the shorter term, the Time Sliding Window technique tends to overestimate the sending rate and

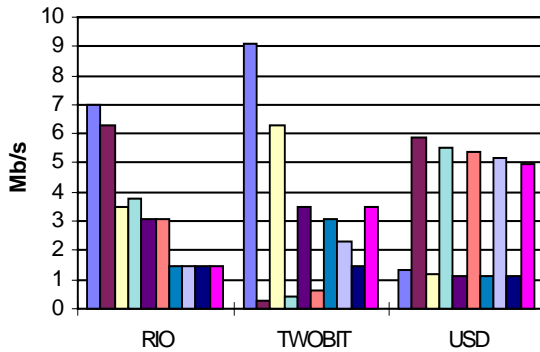


Figure 4: Effect of short-duration traffic with 33Mb/s (1,5)

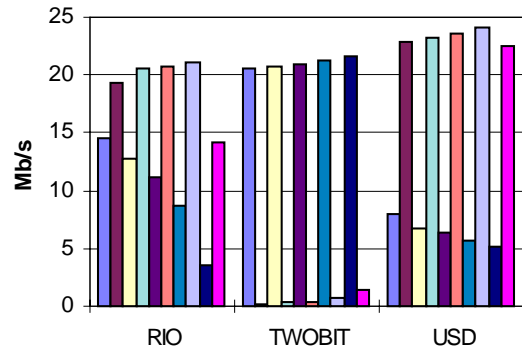


Figure 5 : Effect of short-duration traffic with 155Mb/s (1,5)

mark more packets as out of profile than is necessary. We have performed some experiments to test how sensitive the calculated sending rates are to the length of the time window (the results are not shown here for lack of space), and found that the closer the time windows are to the round trip time, the more accurate the estimates are. This implies that it may be difficult to estimate sending rates accurately in the absence of information about round-trip times, which is the case if the profile meter is de-coupled from the sender. In the 155Mb/s (1,5) case, the bandwidth allocation is more reasonable in the sense that the higher bandwidth sources receive greater bandwidth than the lower bandwidth ones, except for source 10 that gets slightly lower

bandwidth than source 1. However, the bandwidth allocations are not in proportion to the respective expected average bandwidths. We have seen earlier (section 3.1) that the bandwidth at the bottleneck that is in excess of the aggregate expected bandwidth is more or less evenly allocated among all the sources – in this case, where all the traffic flows are of much shorter duration, RED is unable to enforce fair sharing of the excess bandwidth. The higher bandwidth sources, in the short term, are able to grab a higher share of the excess bandwidth since their congestion windows open up faster (due to lower drop rates at the routers that implement RIO) than those of the lower bandwidth sources.

The numbers for the two-bit scheme are interesting. The even sources (i.e., those with expected bandwidths of 5Mb/s), for the most part, show a lower bandwidth allocation compared to the odd sources (i.e. those with expected bandwidths of 1Mb/s) in both the scenarios. This anomalous behavior is caused by the way premium traffic is managed under the two-bit scheme: the monitor at the edge of the network generates premium tokens at a fixed rate, based on the peak bandwidth reserved by the call set up mechanism. Since TCP-Reno starts by doubling the size of the congestion window every round trip time, the traffic generated is bursty and the sending rate can exceed the reserved bandwidth in the short term. As a result, the premium traffic, on reaching the monitor, faces an absence of premium tokens and gets dropped. This happens in the initial stages of the opening up of the congestion window (before *ssthresh* is reached) and fast recovery does not occur because there are not enough lost packets. Thus, TCP times out multiple times and causes the congestion window to open up very slowly. In the long run, this effect is absent because over longer time scales, enough premium tokens are generated to absorb such bursts. Thus, in the short term, the sources with the lower expected bandwidth are able to make use of the excess bandwidth that is freed up by a lack of premium traffic and therefore show higher bandwidth utilization. In the case of the USD scheme, the short term bandwidth allocations match the long term case: the figures in each of the two scenarios show that the bandwidth shares for the sources are in proportion to their expected bandwidth profiles. Thus, for short-lived flows such as web traffic, where end-to-end mechanisms are harder to implement because they need at least one round trip time to react, USD is an attractive option for providing differentiated services.

3.3 Non-Responsive Sources

We used two different kinds of non-responsive sources, CBR (Constant Bit Rate) sources that send at a fixed rate and malicious sources that flood the network by sending as fast as they can. None of these sources back off in the face of congestion. We first investigated the effects of CBR sources on each of the three schemes. The simulations were repeated using bottleneck links of capacity 33Mb/s and 155Mb/s. In the simulations, sources 1 and 2 were CBR sources transmitting at 5Mb/s and the other sources were TCP-Reno sources where the odd numbered sources had an expected bandwidth of 1Mb/s and the even numbered sources had an expected bandwidth of 5Mb/s. All CBR traffic was marked as out of profile.

Figure 6 shows the bandwidth allocation in each of the three schemes when the bottleneck bandwidth is 33Mb/s and Figure 7 shows the same when the bottleneck bandwidth is 155Mb/s. In both figures, the first two bars in each cluster are for the CBR sources and the rest are for the TCP-Reno sources. The numbers for the 33Mb/s (1,5) case indicate that in the case of RIO, the CBR sources are able to get close to their expected bandwidth (i.e., about 5Mb/s) despite all of

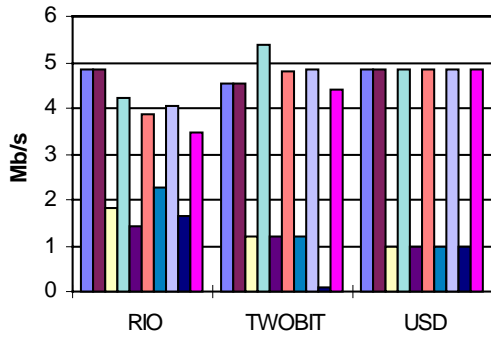


Figure 6: Effect of two non-responsive (CBR) sources with 33Mb/s (1,5)

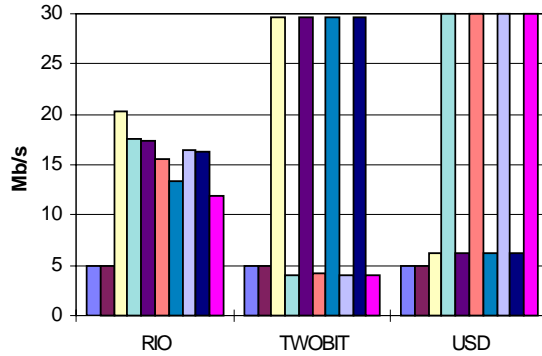


Figure 7: Effect of two non-responsive (CBR) sources with 155Mb/s (1,5)

their traffic being marked out of profile (i.e. lower priority). However, the TCP-Reno sources perform reasonably well, the sources with higher expected bandwidth show a bandwidth share of about 4Mb/s which is close to what the expected bandwidth profile is. The sources with a lower expected bandwidth show bandwidth shares that are higher than their expected bandwidth profiles, closer to about 2Mb/s rather than 1Mb/s. As in section 3.1, we expect that this can be remedied by an appropriate adjustment of the RIO parameters. In the 155Mb/s (1,5) case, the CBR sources get about 5Mb/s as in the previous case. All the other sources (that are all TCP-

Reno sources) receive bandwidths that are in excess of their expected average bandwidths, and it is clear that the excess bandwidth at the bottleneck is allocated fairly evenly among the TCP sources by the RED mechanism, as was the case in Section 3.1.

For the two-bit scheme, the CBR sources get close to their expected bandwidth (about 4.5Mb/s) in both the scenarios. The sources with the higher expected high bandwidth, whose packets are all marked as premium traffic, also get close to 5Mb/s bandwidth (about 4.8Mb/s from the graphs). In the 33Mb/s (1, 5) case, 3 of the 4 sources that have a low expected bandwidth profile of 1Mb/s, get a slightly higher bandwidth allocation than their expected profiles. All of this comes at the expense of the fourth low bandwidth source, which gets almost close to zero bandwidth. Thus, it appears that in the two-bit scheme, if there are non-cooperative, bandwidth hungry sources (even if all their packets are marked out of profile), other well-behaved, low priority traffic may get starved. The situation is slightly different when there is excess bandwidth at the bottleneck, i.e., the 155Mb/s (1,5) case. In this scenario, the premium bandwidth sources do not get any more than their expected bandwidths because premium traffic that is out of profile gets dropped. Consequently, the excess bandwidth at the bottleneck is shared evenly by the low bandwidth TCP-Reno sources, that now show a higher bandwidth allocation than the higher bandwidth sources. Finally, the numbers for the USD scheme show that the bandwidth allocations are commensurate with the expected bandwidth profiles, which implies that the USD scheme is effective in the presence of non co-operative CBR sources.

To investigate the effect of malicious sources, we again used bottleneck bandwidths of 33Mb/s and 155Mb/s. Sources 1 and 2 were set to be malicious sources that declare an expected bandwidth of 1Mb/s and then try to flood the network by sending as fast as they can. The other sources are TCP-Reno sources where the odd numbered sources have an expected bandwidth of 1Mb/s and the even numbered sources have an expected bandwidth of 5Mb/s. Note that these scenarios are different from the CBR cases described earlier, because in the CBR cases, the sources did not try to flood the network in addition to not responding to congestion. The purpose of simulating malicious sources here is to see if they have a negative effect on low priority traffic by filling up the queues allocated for such traffic.

The results for the 33Mb/s (1,5) scenario are shown in Figure 8. In the RIO case, the low bandwidth sources get very little bandwidth – sources 3 and 5 get nothing and sources 7 and 9 get about 0.1Mb/s. The high bandwidth sources do not do well either, none of the sources are able to get more than 2.5Mb/s. The reason for this is that the packets for the malicious sources fill up the queues on the bottleneck routers very quickly and causes packets from other sources to be dropped more often. Consequently, the TCP windows do not open up to the extent required to get the expected average bandwidth. For the two-bit scheme, the low bandwidth sources get very little bandwidth (as in the RIO case) for the same reason and the bulk of the bandwidth goes to the malicious sources. However, in this case, the malicious sources grab about 6Mb/s each as

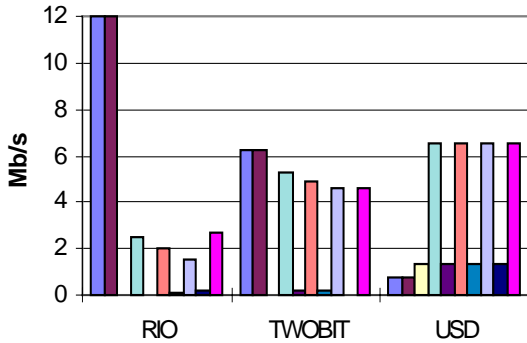


Figure 8: Effect of two malicious sources with 33Mb/s (1,5)

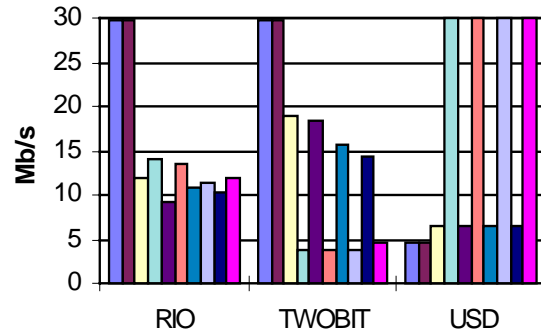


Figure 9: Effect of two malicious sources with 155Mb/s (1,5)

opposed to 12Mb/s each in the RIO case. The reason is that the packets from the high bandwidth sources in the two-bit scheme are enqueued in a separate, higher priority, premium traffic queue. As a result, the high bandwidth sources are able to get close to their expected bandwidths of 5Mb/s while the low bandwidth sources get very little – sources 3, 5 and 7 get less than 0.2Mb/s and source 9 gets nothing. Thus, in both RIO and two-bit schemes, a malicious source can easily deprive low priority traffic of their fair share, but the high priority traffic tends to do better in the two-bit scheme because such traffic is enqueued in a separate, high priority queue. In the case of the USD scheme, the malicious sources are penalized and actually receive slightly less than their average expected bandwidth (0.79Mb/s in place of 1Mb/s) while the other sources get slightly more than their expected bandwidth – the low bandwidth sources get 1.31Mb/s instead of 1Mb/s and the high bandwidth sources get 6.54Mb/s instead of 5Mb/s.

The results for the 155Mb/s (1,5) scenario are shown in Figure 9. In this scenario, the malicious sources for the RIO and the two-bit schemes are able to grab almost the full bandwidth of their

outgoing links, which is set to 30Mb/s in the simulation. In the RIO case, the TCP-Reno sources are able to get more than their expected bandwidth profiles, but not in proportion to their expected bandwidths. Again, the reason is that the excess bandwidth (that is consumed only by out of profile packets) is more or less evenly distributed by the RED scheme among all the sources. In the two-bit case, the premium sources are able to get slightly less than their expected bandwidths (in the range 3.5 – 4Mb/s instead of 5Mb/s). This is somewhat surprising since, given that the bottleneck bandwidth in this case is 155Mb/s, and the fact that premium traffic gets higher priority than best effort traffic, one would have expected that the premium traffic would get the full 5Mb/s bandwidth. We conjecture that this is because the malicious sources constantly fill up the lower priority queues and therefore the routers spend more time in processing the lower priority queues than the higher priority ones. Consequently, the congestion windows for the higher bandwidth premium traffic does not open up fast enough to utilize the requisite expected bandwidth. The excess bottleneck bandwidth is more or less evenly allocated only among the low bandwidth sources because any premium traffic that is out of profile is dropped. This causes the low bandwidth sources to get a higher bandwidth allocation than the high bandwidth sources. Lastly, the USD scheme again penalizes the malicious sources and the allocation of bandwidth among the well behaved TCP-Reno sources is in proportion to their expected average bandwidths – the figures show that the low bandwidth sources get about 6Mb/s while the high bandwidth sources get about 30Mb/s.

The performance of the three schemes indicate that in the presence of non-responsive sources, low priority traffic could get starved when the RIO or two-bit schemes are used. The two-bit scheme ensures that the high bandwidth sources (which send premium traffic) are able to utilize their average profile bandwidth whereas the RIO scheme is not as effective in restricting non-responsive sources (especially malicious ones) from grabbing as much bandwidth as possible, even at the cost of high priority traffic. It is necessary to develop stricter monitoring mechanisms for the RIO and two-bit schemes in order to identify and penalize non responsive traffic flows (for example, the TCP-friendly approach outlined in [16]) for such schemes to be more effective. The USD scheme performs well in that it is able to effectively penalize non-responsive and malicious sources, but at the cost of increased complexity in the core routers.

3.4 Flows with High Expected Bandwidth

Another set of simulations was performed to investigate the effects of over allocation of bandwidth. In this case, we set the expected bandwidth of all the sources to 5Mb/s and used a 33Mb/s bottleneck link. The performance is shown in Figure 10. For the RIO scheme, the bandwidth allocation is fairly even, which indicates that the RED scheme is nearly fair if all connections have the same weight. In the case of the two-bit scheme, only 6 of the 10 connections are able to get any bandwidth, the other connection requests are denied at set up time. The graphs indicate that the monitoring scheme for the premium packets is effective in ensuring that the allowed connections stay within their profiles. The graphs for the USD scheme show that the bandwidth allocation is equal for each of the 10 connections, which is to be expected. The lesson here is that although the RIO and the USD schemes adapt well in the case when the network is under provisioned and all connections have equal weights, the two-bit scheme can handle such a situation only by denying service to some of the connections.

In a setting where organizations or individuals have already paid an ISP for service guarantees, this may not be a feasible option. Hence, in the two-bit case, networks perforce have to be over provisioned, a solution that under-utilizes the network and may not be economically sustainable in the long run.

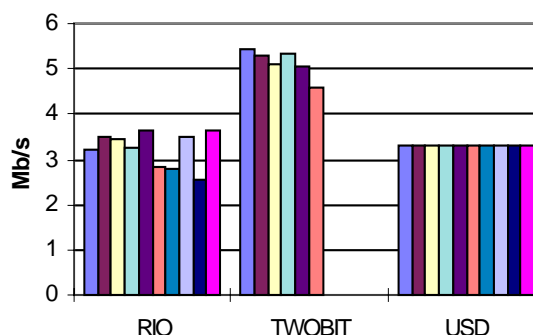


Figure 10: Effect of high bandwidth flows

4. Summary

In this paper, we have evaluated the bandwidth allocation performance of the RIO, the two-bit and the USD schemes. We have shown that matching the network bottleneck bandwidth to the aggregated expected bandwidth profiles over all the traffic flows through the bottleneck is an important issue. While the RIO and the two-bit schemes perform well when there is no mismatch, such is not the case if the bottleneck bandwidth available is in excess of the aggregated expected bandwidth through the bottleneck. This does not mean that the sources get less than their

expected bandwidths, it is just that the bandwidth allocation in such a situation is not commensurate with the user expectations. On the other hand, the USD scheme with PFS is able to do a fair allocation of bandwidth that is in keeping with user expectations. The issue of short term flows also needs to be addressed in view of the high percentage of web traffic on the Internet – the RIO and two-bit schemes need to be modified in some way to react more quickly to changes in network conditions to make this possible. The USD scheme is able to achieve this since it does not rely on end-to-end congestion control schemes to enforce fairness. In the case of non-responsive sources, RIO performs better than the two-bit scheme when the non-responsive sources are CBR, however, malicious sources that flood the network are able to starve low priority traffic. The USD scheme is able to provide traffic isolation under such circumstances. For high bandwidth flows, all the schemes perform well, except that in the two-bit scheme case, an over-committed network may deprive some of the sources from getting any bandwidth, instead of allocating the available bandwidth evenly among all active sources. It should be emphasized here that the advantages of USD come at a cost – the cost of implementing PFS in the network core. However, as explained earlier in section 2.3, it is possible to use certain variants of weighted fair queuing [7,13] to provide PFS without incurring great computation complexity – the recently announced PacketStar router from Bell Labs[10] is a case in point.

5. Discussion

The differentiated services model tries to provide quantitative bandwidth (and/or delay) guarantees to users by relying on traffic policing at the network edges and at the same time, avoids any form of signaling to provision the network. Although the two-bit scheme proposes bandwidth brokers that are used off-line to allocate bandwidth (and thus prevent the network from being over-committed), the assumption is that with admission control at the edges of the network, it is possible to limit the total traffic for a particular user (or class). However, traffic flows inside the network tend to be dynamic with respect to bandwidth demands as well as source/destination pairs. In order to perform proper admission control, it is necessary for the edge network nodes to have global knowledge about all the bottleneck links inside the network, which is difficult to achieve in practice.

Alternatively, the network has to provision enough bandwidth for the worst-case traffic distribution. However, over-provisioning can lead to low utilization of network resources, and in any case, it may not be easy to estimate worst-case traffic distribution in the long run. In addition, in a large network, network wide over-provisioning may not be economically feasible. The USD scheme addresses the problem by making the guarantee relative. Since the bandwidth is allocated according to the weights assigned to each user, the USD scheme can adopt automatically to multiple situations. Under normal circumstances, we expect a bandwidth lookup table consisting of user's network prefix and its associated share to be distributed via a network management system as part of the provisioning.

Although the three schemes described in this paper address the issue of policing the network, none of them effectively solve the problem of network provisioning, with the possible exception of the two-bit scheme that advocates use of off-line bandwidth brokers (discussed in the previous paragraphs). We believe that some form of coarse grained signaling for aggregated flows will be necessary for this purpose in conjunction with a policing mechanism that enforces provisioning guarantees. We shall refer to such a protocol as the Trunk Provisioning Protocol, or TPP. TPP can be built with a revised RSVP or as a new, light-weight protocol. The main purpose of TPP is to provide a mechanism for provisioning network trunks between ingress and egress routers on the network. As it only deals with aggregated flows, the operation of TPP is de-coupled with setting up and tearing down individual microflows. The provisioning of trunks needs to be adjusted only when there are significant and persistent changes among the aggregated flows. Thus, the overheads of TPP are likely to be very low.

The Trunk Provisioning Protocol can be designed to provide either explicit or proportional provisioning and can be used in conjunction with USD for provisioning different shares along different paths. For example, if a customer has a subsidiary in another country, it can have a higher share on the international path to that country and a lower share on a path to a neighboring ISP. For the RIO or the two-bit schemes, TPP can be used to set up egress-specific guarantees. For example, instead of providing a fixed amount of premium bandwidth anywhere in the network, one can guarantee different bandwidths between the access point of a customer to specific egress points in the network. Working in this mode, TPP is more like a trunk reservation protocol. If sufficient bandwidth on a specific path is not available, the provisioning will fail. This has the advantage that while it might not be possible to guarantee a given bandwidth to all

destinations in the network, it might still be possible to guarantee bandwidths for a specific destination that the user might be interested in at the moment. The specification and implementation of such a protocol needs to explore complex issues involving network provisioning based on long-term traffic characteristics and is left as future work.

Acknowledgments

The authors would like to thank Srinivasan Keshav for suggesting some of the ideas that eventually led to the work described in this paper.

References

- [1] D. Black, S. Blake, M. Carlson, E. Davies, Z. Wang, and W. Weiss. An Architecture for Differentiated Services. Internet Draft, May 1998.
- [2] D. Clark and W. Fang. Explicit Allocation of Best Effort Packet Delivery Service. Draft, 1998.
- [3] D. Clark, S. Shenker, and L. Zhang. Supporting Real-Time Applications in an Integrated Services Packet Network: Architecture and Mechanisms. *In Proceedings of ACM SIGCOMM'92*, August 1992.
- [4] D. Clark, J. Wroclawski. An Approach to Service Allocation in the Internet. Internet Draft, July 1997.
- [5] M. Degermark, A. Brodnok, S. Carlsson, and S. Pink. Small Forwarding Tables for Fast Routing Lookups. *In Proceedings of ACM SIGCOMM'97*, Cannes, France, September 1997.
- [6] K. Fall and S. Floyd. Promoting the Use of End-to-End Congestion Control in the Internet. Submitted for publication, February 1998.
- [7] S. Golestani. A self-clocked fair queuing scheme for broadband applications. *In Proceedings of IEEE INFOCOM'94*, pages 636-646, April 1994.
- [8] V. Jacobson. Differentiated Services Architecture. Talk in the Int-Serv WG at the Munich IETF, August 1997.
- [9] S. Keshav. REAL 5.0 Manuals. <http://www.cs.cornell.edu/skeshav/real/man.html>, August 1997.

- [10] V. P. Kumar, T. V. Lakshman, and D. Stiliadis. Beyond Best Effort: Router Architectures for the Differentiated Services of Tomorrow's Internet. *IEEE Communications Magazine*, May 1998.
- [11] K. Nichols, V. Jacobson, and L. Zhang. A Two-bit Differentiated Services Architecture for the Internet. Internet Draft, November 1997.
- [12] A. K. Parekh and R. G. Gallager, "A Generalized Processor Sharing Approach to Flow Control - the Single Node Case. *In Proceedings of IEEE INFOCOM'92*, May 1992.
- [13] M. Shreedhar and G. Varghese. Efficient Fair Queuing Using Deficit Round Robin. *In Proceedings of ACM SIGCOMM'95*, Cambridge, Massachusetts, August 1995.
- [14] K. Thompson, G. Miller, and R. Wilder. Wide-Area Internet Traffic Patterns and Characteristics. *IEEE Network*, November 1997.
- [15] Z. Wang. User-Share Differentiation – Scalable Service Allocation for the Internet. Internet Draft, November 1997.
- [16] Z. Wang. USD: Scalable Bandwidth Allocation for the Internet. *In Proceedings of HPN'98*, September 1998 (to appear).
- [17] L. Zheng, S. Deering, S. Estrin, S. Shenker, and D. Zappala. A New Resource Reservation Protocol. *IEEE Network*, September 1993.