Diffserv Working Group
Internet Draft
Expires: August 2001
Document: <draft-jagannathan-diffserv-pf-00.txt>
Category: Informational

P. Jagannathan
OSU
A. Durresi
OSU
Raj Jain

Raj Jain is now at Washington University in Saint Louis, jain@cse.wustl.edu http://www.cse.wustl.edu/~jain/

February 2001

Proportional Forwarding PHB

Status of this Memo

A PDF version of this document is available upon request.

1. Abstract

We propose a new Per Hop Behavior (PHB) for differentiated services in IP networks. In this PHB, each network subscriber (or flow) receives a bandwidth allocation proportional to the subscribed information rate. This new PHB is called "Proportional Forwarding PHB" or PF PHB. The marking mechanism at the source or ingress router and the queuing and discard behavior at core routers are clearly described.

Three different sample marking algorithms are proposed in this document and are analyzed using simulations. It is shown that it is possible to obtain proportional bandwidth allocation using proper marking mechanism at the ingress routers and multi-level threshold-based dropping mechanism at the core routers.

2. Conventions used in this document

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC-2119.

3. Introduction

In the current Internet, users might have different service requirements. The most important service parameter - the subscribed information rate (SIR) - varies widely from one user to another. However, in the best effort service of the Internet, SIR is not used to allocate bandwidth during congestion. A user with a higher SIR generally pays more than a user with a lower SIR. Therefore, during congestion, a user with a higher SIR will expect to be allocated more bandwidth than the bandwidth allocated to a user with a lower SIR. We define a new method of bandwidth allocation, called Proportional Allocation of Bandwidth (PAB), in which bandwidth must be allocated in proportion to the SIR of the competing flows.

In this document, we propose a PHB, the Proportional Forwarding PHB which achieves this proportional allocation of bandwidth. A labeling mechanism at the ingress router encodes the ratio of source’s sending rate to its subscribed information rate to set the Differentiated Services Code Point (DSCP). In the core routers, a multi-level threshold-based active queue management scheme is used to drop packets.

We avoid the storing of per-flow state information by encoding the ratio of a flow’s data rate to its SIR in the form of a DSCP on its packets. In the interior of the network, the routers use these DSCPs for differentiating between packets during congestion. As, all the marking is done at either the source or the first network element – ingress router after the source, only the marking device needs to have information about the flow’s SIR. The core router drops packets based on the DSCPs and the current level of congestion in the router. Thus, no state information is stored in the core of the network.

4. Purpose

The PF PHB achieves differentiation among flows based on their SIR. The significant advantage of PF PHB is that there is no state information stored in the network. When there are numerous flows having small SIR and service differentiation has to be done based on their SIR and it is not scalable to keep state information for each of them, PF PHB can be used.

5. Proportional Allocation of Bandwidth - PAB

The principle behind PAB is that the allocation of bandwidth should be in proportion to SIR of the flows sharing the link. The SIR of a flow is one of the most important service parameters for the flow.

It is, therefore, important to consider both the flow's data rate and its SIR, to allocate bandwidth.

According to the definition of max-min fairness as defined in [1], each flow is allocated bandwidth as given by:

```
Alloc (i) = Min{send(i), rr}
Sum_i Alloc (i) < Available Bandwidth
```

Here send(i) is the data rate of the ith flow and rr is the maximum rate that satisfies the above inequality. All flows sending at a rate less than rr are allocated bandwidth equal to their sending rate. Flows sending more than rr have their throughput reduced to rr. The main problem with this allocation is that the SIR of the flow is not considered in bandwidth allocation.

Given that users pay proportional to their SIR, the bandwidth should be allocated proportional to SIRs. Of course, this requirement must be satisfied with full network utilization. Therefore in PAB the allocation of bandwidth is given by:

```
Alloc(i) = Min{ send(i), frac * SIR(i) }
Sum_i Alloc(i) < Available Bandwidth
```

Here, SIR(i) is the SIR of the ith flow and "frac" is the maximum fractional multiplier (between 0 and 1) that satisfies the above inequality.

If the data rate of a flow is below its allowed throughput "frac"*SIR then it does not suffer any packet loss. Further, if a flow has a data rate less than its allowed fraction of SIR, then the remaining excess bandwidth is also shared among other flows in proportion to their SIRs. No flow is allowed to send more than its SIR during congestion. The throughput of any flow sending more than the allowed fraction of SIR is reduced to its fair allocation. Thus PAB differentiates between flows and allocates bandwidth in proportion to the SIR of the flows.

6. Packet Marking at the Edge of Network

The granularity of bandwidth allocation depends upon the number of DSCPs used to mark the packets. A larger number of code points allows more granular allocation in the sense that the actual allocation and desired allocations (given by the above PAB formula) are very close. The challenge is to achieve actual allocations as close as possible to the desired allocations with a limited number of code points. We explore several potential marking mechanisms in this section.

Packets can be marked at the source or the ingress router. The marking device has a knowledge of SIRs of all the flows through it. A flow's packets are marked with different DSCPs depending on the ratio of the flow rate (FR) to the SIR. The total number of DSCPs is fixed for all flows, but the number of DSCPs used at any time for a flow depends on its FR. As the ratio of FR to SIR increases for a flow, more and more packets will be marked with "lower priority" DSCPs.

Given n code points, we use n token buckets such that a packet can consume tokens only from one token bucket. The user should be able to send data at or below its SIR. The average data rate should not exceed the user's SIR. So the sum of the token rates of all the token buckets must be equal to the SIR of the flow. The SIR however is distributed among all the token buckets. Therefore, the token rate of an individual token bucket is a fraction of the flow's SIR. This fraction is equal to the fraction associated with the DSCP corresponding to that token bucket. The sum of the fractions associated with all the DSCPs is 1. Thus, the token rate is the product of the fraction of its DSCP and the SIR of the flow and the sum of the token rates is the SIR of the flow:

Token Rate of jth bucket = Frac_j * SIR
Sum_j_ Frac_j = 1

The significance of the value of the fractions is discussed later. The flow rate determines the actual DSCPs that the packets get. As the ratio of flow rate to SIR increases, more and more packets will be marked with lower priority DSCPs. The token bucket size allows for bursts in the flow rate. However, the long term rate of the flow can never exceed its SIR.

Alternatively, the flow's packets may be marked with the lowest priority if the flow exceeds its SIR for a long period of time. However a uniform methodology of marking packets for all flows should be used when SIR is exceeded for a longer period than the allowed burst.

7. Packet Dropping Mechanism at the Core Routers

At the core routers, packets are dropped based on their DSCPs. An active queue management scheme similar to random early drop (RED) can be used for dropping packets during congestion. Normally, RED implementations have three parameters: Qmin, Qmax, and Pmax. The probability of drop is a linear function of average queue length. The probability is zero if the average queue length is less than Qmin. The probability is Pmax if the average length is more than Qmax. The probability increases linearly between 0 and Pmax as the average queue length increases from Qmin to Qmax. Extending this scheme to n classes of packets, we get an n-level RED (herein called n-RED). Packets with ith code point use RED with Qmin_i, Qmax_i, and Pmax_i. Since all packets of the flow stand in the same queue, the total average queue length is compared with the ith threshold to decide whether a packet with the ith DSCP should be dropped on arrival. As shown in Figure 1, code point i has a "lower priority" than code point i-1 in the sense that for any given queue length, packets with ith DSCP have a higher drop probability than those with code point i-1. Note that we have used only one average queue length. This is simpler than the other possible alternative of keeping a count of packets of each DSCP in the queue.

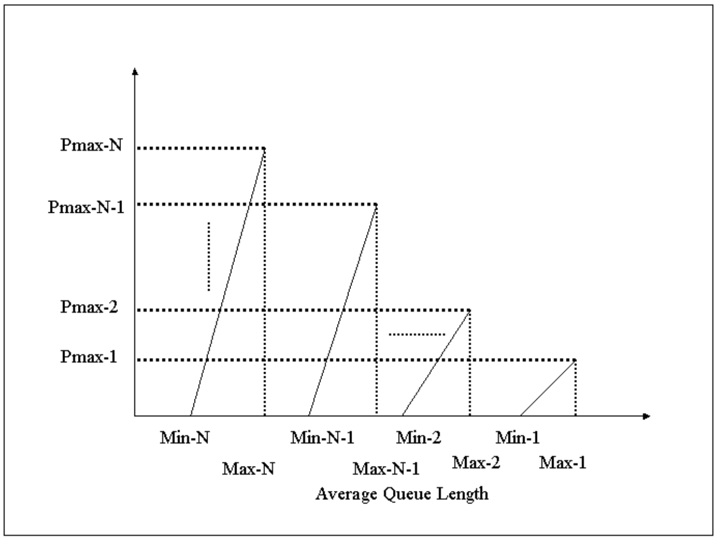

Figure 1 Multilevel Threshold based Packet Handling Mechanism

8. Determination of the DSCP fractions

If n DSCPs are available for PF PHB, we require that all marking devices should use the same number (n) of DSCPs and the same algorithm for marking packets. We have studied three different marking algorithms. Each of these algorithms uses n token buckets with the token rate of the jth bucket is frac_j times SIR of the flow. The three algorithms are different in the way fractions are set. The three algorithms are:

a. Fractions with equal Value - Equal fractions
b. Fractions forming arithmetic progression - AP fractions
c. Fractions forming geometric progression - GP fractions

8.1. Equal Fractions

All the fractions are of equal value. So, if there are n DSCPs, then each DSCP has the fraction value frac_i set to 1/n. So the sum of fractions is 1. For example, with 8 DSCPs the fractions associated are all 1/8.

8.2. Arithmetic Progression (AP) Fractions

The fractions form an arithmetic progression. Unlike equal fractions, the values for AP fractions do not have the same value. The n fractions determining the token rates of the n token buckets are:

a, a+d, a+2d, …, a+(n-1)d

For simplification, we assume "a" to be equal to "d". So this gives the values of the fractions to be

d,2d,3d, …, nd

Since the sum of the arithmetic progression has to be equal to 1, we obtain the following value for "d",

d = 2 / (n * (n+1))

For example, when there are 8 DSCPs, the values of fractions are 1/36, 2/36, 3/36, 4/36, 5/36, 6/36, 7/36, and 8/36, in decreasing order of priority.

8.3. GP Fractions

The fractions form a geometric progression. Similar to arithmetic progression, the values of the fractions are assigned such that higher priority DSCP IDs are associated with smaller values in the geometric progression. So the fractions are:

a, ar, ar^2, .. , ar^(N-1)

Again as in arithmetic progression, to simplify calculations, we assume "a" to be equal to "r". So the values of the fractions are:

r, r^2, r^3, .. ,r^N

Since the sum of the geometric progression has to be equal to 1, the value for r is given by:

r (2 –(r^N)) = 1

For n = 8, r gets a value approximately equal to 1/2. So the fractions have the following values 1/2, 1/4, 1/8, 1/16, 1/32, 1/64, 1/128 and 1/128 in increasing order of priority. The last two fractions are made equal so that the sum of the fractions is 1.

9. Summary of Analysis

Detailed results of simulation analysis using the above three marking algorithms are presented in the Appendix. The results show that all three methods of marking perform significantly better than no marking. Equal fraction does not perform as well as the other two under severe congestion.

In conclusion, we have proposed a PHB in which subscribers get bandwidth allocations proportional to their subscribed rate and have shown that such a behavior is achievable using DSCP marking. We have proposed three different sample marking algorithms and have presented detailed simulation analysis of these algorithms.

10. Issues

10.1. Recommended PHB Ids

According to [9], the PHB Ids that should be used must be 8 values from (xxxxxxxxxxxx0011) which have been reserved for experimental and local use.

10.2. Example Uses

Consider a case where an ISP wants to provide many different services as needed by customers. There are no guarantees provided. However, customers are assured that they will get a better bandwidth allocation than customers who have subscribed information rates below their SIR. Further this scheme can also be used to achieve proper service agreements among many Internet content providers whose flows traverse the same congested link, irrespective of the sensitivity of their flows to congestion.

10.3. Security Considerations

There are no known security considerations, other than the DS domain should be able to limit the traffic entering the domain according to the subscribed profiles.

11. References

[1] I. Stoica, S. Shenker and H. Zhang, "Core-Stateless Fair Queuing : Achieving Approximately Fair Bandwidth Allocations in High Speed Networks," Computer Communication Review. vol.28, no.4 Oct. 1998. p. 118-30.

[2] A. Demers, S. Keshav and S.Shenker, "Analysis and Simulation of a Fair Queuing Algorithm," Computer Communication Review SIGCOMM '89 Symposium Communications Architectures and Protocols v 19 n 4 1989 p.1-12.

[3] D. D. Clark and W. Fang, "Explicit allocation of Best-Effort Packet Delivery Service," IEEE/ACM Transactions on Networking, v 6 n 4 August 1998 p.362-73.

[4] S. Floyd and V. Jacobson, "Random Early Detection Gateways for Congestion Avoidance," IEEE/ACM Transactions on Networking, v 1 n 4 August 1993 p.397-413.

[5] M. Shreedhar and G. Varghese, "Efficient Fair Queuing using Deficit Round Robin," Communication Review ACM SIGCOMM '95 v 25 n 4 1995 p.231-42.

[6] V. Jacobson, "Congestion Avoidance and Control," Computer Communication Review SIGCOMM '88 Symposium: Communications Architectures and Protocols v 18 n 4 1988 p.314-29.

[7] B. Braden et al. "Recommendations on Queue Management and Congestion Avoidance in the Internet," RFC 2309, April 1998.

[8] S. Blake et al. "An Architecture for Differentiated Services," RFC 2475, December 1998.

[9] S. Brim, B. Carpenter, F. Le Faucheur, "Per Hop Behavior Identification Codes," RFC 2836, May 2000.

[10] J. Heinanen, R. Guerin, "A Single Rate Three Color Marker," RFC 2697, September 1999.

[11] J. Heinanen, R. Guerin, "A Two Rate Three Color Marker," RFC 2698, September 1999

[12] Z. Cao, Z. Wang and E. Zegura, "Rainbow Fair Queuing: Fair Bandwidth Sharing Without Per-Flow State," Proceedings IEEE INFOCOM 2000,Conference on Computer Communications, p.922-31 vol.2.

[13] J. Heinanen, F.Baker, W.Weiss, J.Wroclawski, "Assured Forwarding PHB Group," RFC 2597. June 1999.

[14] S. Floyd, K. Fall, "Promoting the Use of End-to-End Congestion Control in the Internet," IEEE/ACM Transactions on Networking, v 7, n 4. 1999, p 458-472.

[15] S. Floyd, V. Jacobson, "Link Sharing and Resource Management Models for Packet Networks," IEEE/ACM Transactions on Networking. v 3, n 4. August 1995, p 365-386.

[16] M. Goyal, A.Durresi, R.Jain and C.Liu, "Performance Analysis of Assured Forwarding," IETF Draft.- draft-goyal-diffserv-afstdy-00 - , work in progress, February 2000.

[17] R. Jain, "The Art of Computer Systems Performance Analysis," John Wiley and Sons Inc., 1991.

[18] D. Lin and R. Morris, "Dynamics of Random Early Detection," Computer Communication Review, ACM SIGCOMM 97 Conference. v 27 n 4 1997 p.127-37.

[19] B. Suter, T. V. Lakshman, D. Stiliadis, A. Choudhury , "Efficient Active Queue Management for Internet Routers," Proceedings of INTEROP' 98.

[20] Network Simulator – NS – Available at http://www-mash.cs.berkeley.edu/ns

[21] DiffServ patch for NS, S. Murphy, - Available at http://www.teltec.dcu.ie/~murphys/

12. Author's Addresses

Prasanna K. Jagannathan
Department of CIS,
The Ohio State University,
2015, Neil Ave.,
Columbus, OH 43210

Phone: 614-292-7344
Email: pjaganna@cis.ohio-state.edu

Arjan Durresi
Department of CIS,
The Ohio State University,
2015, Neil Ave.,
Columbus, OH 43210

Phone: 614-688-5610
Email: durresi@cis.ohio-state.edu

Raj Jain
Chief Technology Officer,
Nayna Networks, Inc.
157 Topaz St. Milpitas, CA 95035

Phone: 408-956-8000X309
Email: jain@nayna.com

A. Appendix: Simulation Results

We compared the performance of our technique to implement PAB with equal fractions, AP fractions and GP fractions to the performance of the scheme without any differentiated marking. We used a simulation model and studied the performance in single congested link and multiple congested link configurations. The results of these simulations are presented in this Appendix.

A.1 Single Congested Link

We used the ns-2 simulator [20] for performing simulations. The packets are marked with DSCPs at the edge of the network at ingress routers. When there is congestion, the core routers uses DSCP marking in the packets for dropping the packets.

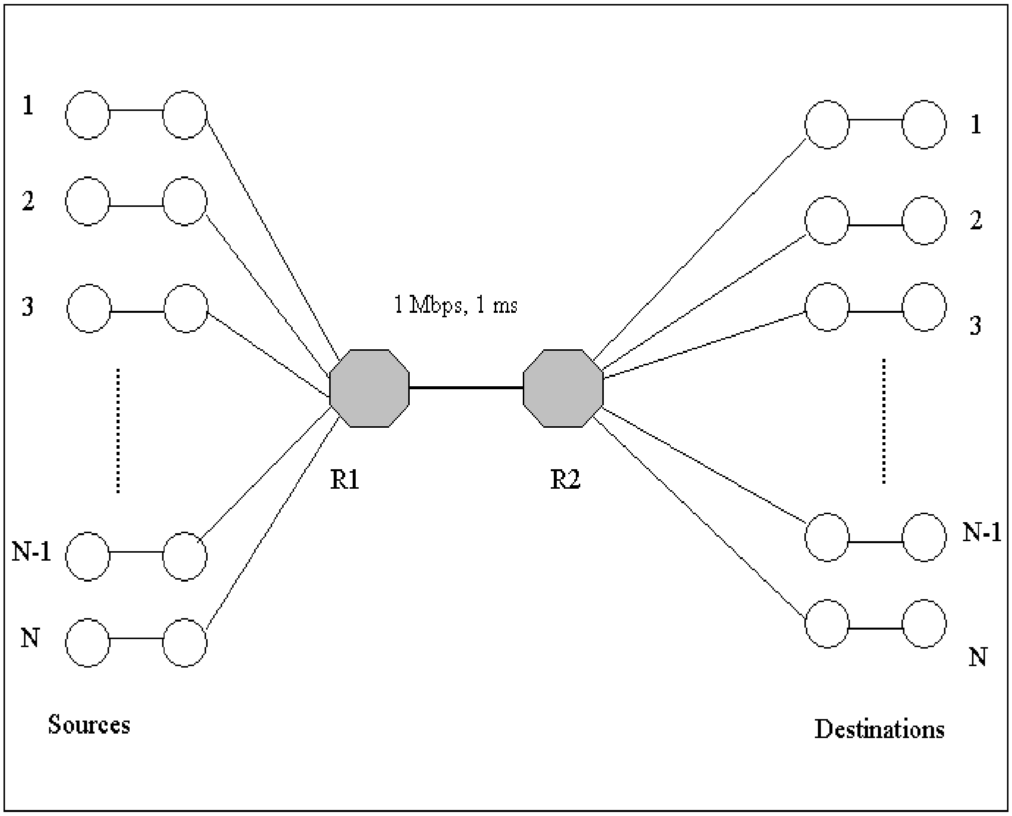

Figure A.1 Single Congested Link - Network Configuration

In the case of a single congested link, the network configuration is as shown in Figure A.1 There are N flows sharing a single congested bottleneck link. The bottleneck link is at the core of the network. The SIR of the flow i was set at i/N*500 kbps. The number of flows sharing the link varied from 3 to 32. The capacity of the bottleneck link is 1 Mbps and the link delay is 1 ms. The capacity of the link buffer was 100 packets. The packet size of the TCP flows was set at 1000 bytes. The packet size of the CBR flows was set at 210 bytes. The parameters for n-RED (n-level drop algorithm) at the core router are shown in Table A.1.

For the token buckets at the ingress routers, the bucket size was fixed at 80,000 bytes. The token rate for each DSCP was determined based on the method of fractions chosen. The token rate for each of the eight DSCPs is the product of the fractions associated with that DSCP and the SIR of the flow.

Table A.1 Parameters for n-RED Packet Dropping Mechanism at the Core Routers

DSCP	Minimum Threshold Qmin	Maximum Threshold Qmax	Maximum Drop Probability Pmax
1 (highest priority)	80	90	1/50
2	70	80	1/45
3	60	70	1/40
4	50	60	1/35
5	40	50	1/30
6	30	40	1/25
7	20	30	1/20
8 (lowest priority)	10	20	1/15

By definition of PAB, each flow should get a share of bandwidth proportional to its SIR. The measure that we used to calculate the effectiveness of the allocation of bandwidth is obtained as follows. The throughput ratio of jth flow [TR(j)] is defined as the ratio of throughput of jth flow to the sum of the throughputs of all flows going through the same link:

TR(j) = Throughput of flow (j) / Sum(Throughput of all flows)

The SIR ratio for the jth flow [SR(j)] is defined as the ratio of SIR of jth flow to the sum of the SIR’s of all flows going through the same link:

SR(j) = SIR of flow(j) / Sum (SIR of all flows)

The allocation ratio for the jth flow [AR(j)] is defined as the ratio of TR(j) to SR(j):

AR(j) = TR(j) / SR (j)

The proportionality index is then computed as follows:

Proportionality Index = [Sum(AR(j))] ^ 2 / (N * Sum([AR(j)] ^ 2))

See [17] for discussions on this formula, which is recommended there as a measure of fairness. In the first set of experiments, all N flows were UDP flows. All UDP flows were also constant bit rate (CBR) flows. The sources were sending data randomly between 10% to 200% of their SIR. The experiments were performed using equal fractions, AP fractions ,GP fractions with n-RED dropping mechanism at the core.

For each marking method, 30 simulations were performed, with the number of flows through the single congested link increasing from 3 to 32. For each configuration, we also simulated the case of a simple RED without any differentiated marking. Figure A.2 shows the performance of the three methods and simple RED.

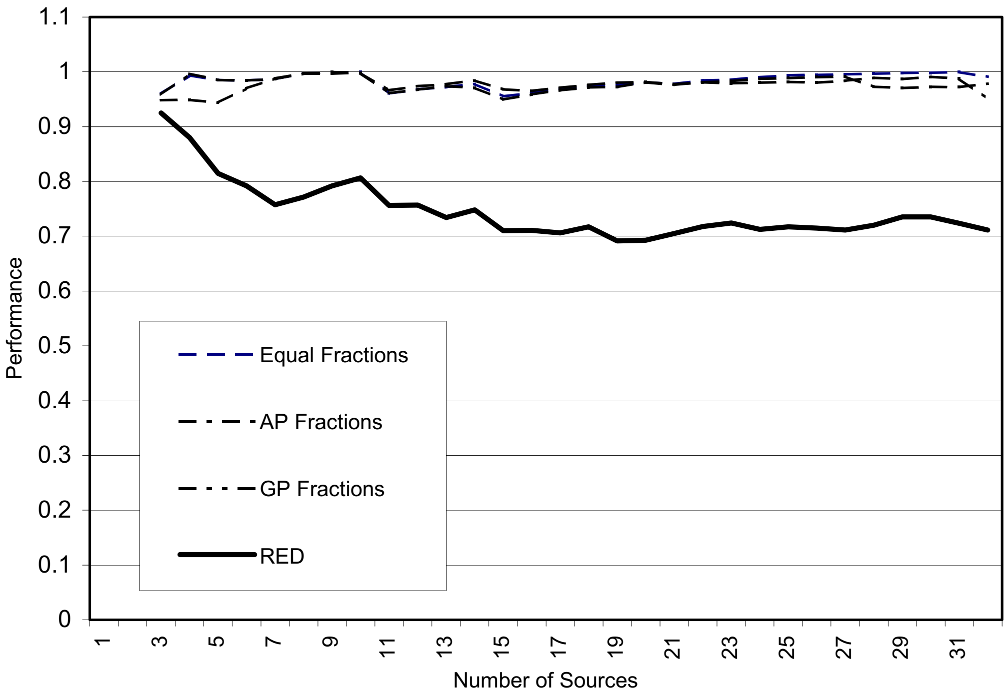

Figure A.2: Performance in Single Congested Link with UDP flows

From Figure A.2 it can be observed that as the number of flows increases, the performance of the three types of fractions are almost similar. However, for simple RED, the performance drastically drops as the number of flows increases. This is due to the fact that RED has no knowledge of SIR and thus cannot differentiate between the flows based on their SIR and thus the bandwidth allocation by RED does not follow the principles of Proportional Allocation of Bandwidth.

In the second set of experiments all N flows were TCP flows, which were Telnet applications with their peak rates set at 400 kbps. As before, the number of flows was increased from 3 to 32 and the three types of fractions and simple RED were used to perform simulations. Figure A.3 shows the performance for TCP flows.

In the case of TCP flows, as the number of flows increases the performance of the three fractions are very good. The TCP flows are congestion sensitive and when there is congestion, the TCP flows tend to share the bandwidth equally among the flows. Our technique achieves proportional bandwidth sharing by using DSCPs and thus achieves good performance. In the case of simple RED, the performance has become much worse than that with UDP flows. During congestion the TCP flows reduce the sending data rate so that the rate of all flows are equal and RED cannot distinguish between flows and thus has poor performance.

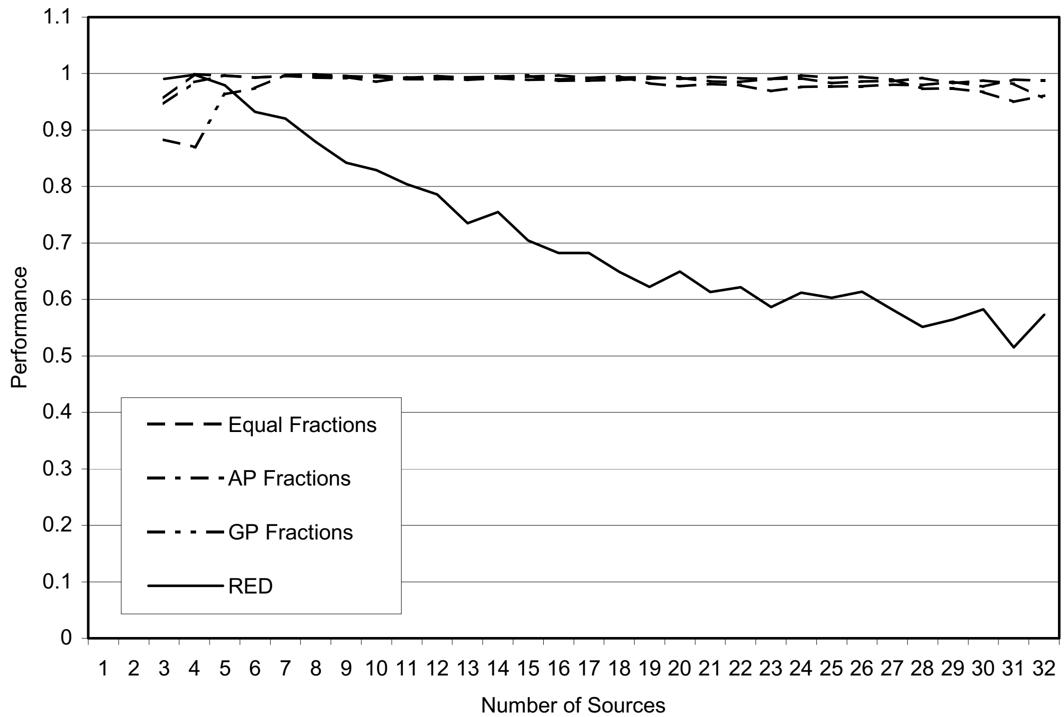

Figure A.3: Performance in Single Congested Link with TCP flows

In the third set of experiments the N flows were mixed. The flows 0,2,4,.. were TCP Telnet flows with the same parameters as before and the flows 1,3,5,.. were CBR UDP flows with the same data rate as before. Again the number of flows was varied from 3 to 32 and the experiments were done for all three types of fractions and simple RED.

Figure A.4 shows the performance with mixed TCP and UDP flows. UDP is congestion insensitive and TCP is congestion sensitive. So UDP flows try to get all the bandwidth and TCP flows get very little. Our technique provides good protection for TCP flows from UDP flows and achieves excellent performance. However, without marking, the TCP flows are not protected and the performance is poor.

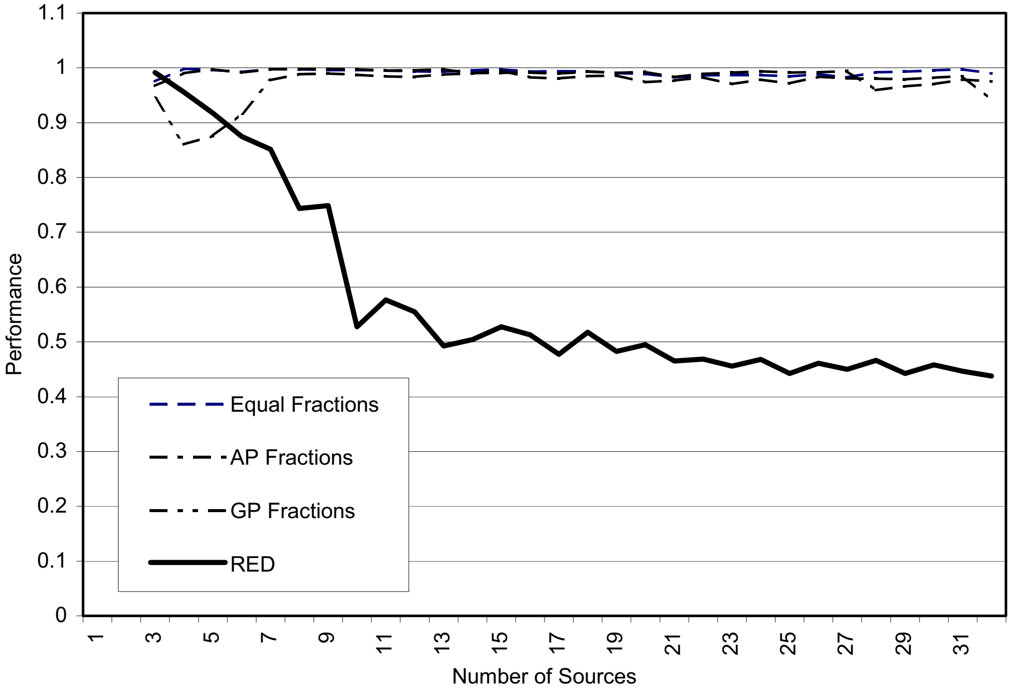

Figure A.4: Performance in Single Congested Link with TCP and UDP flows

A.2 Multiple Congested Links

We used the so called "Parking Lot" configuration to analyze performance when there are multiple congested links. This configuration, shown in Figure A.5, is a typical path taken by a flow in the current Internet.

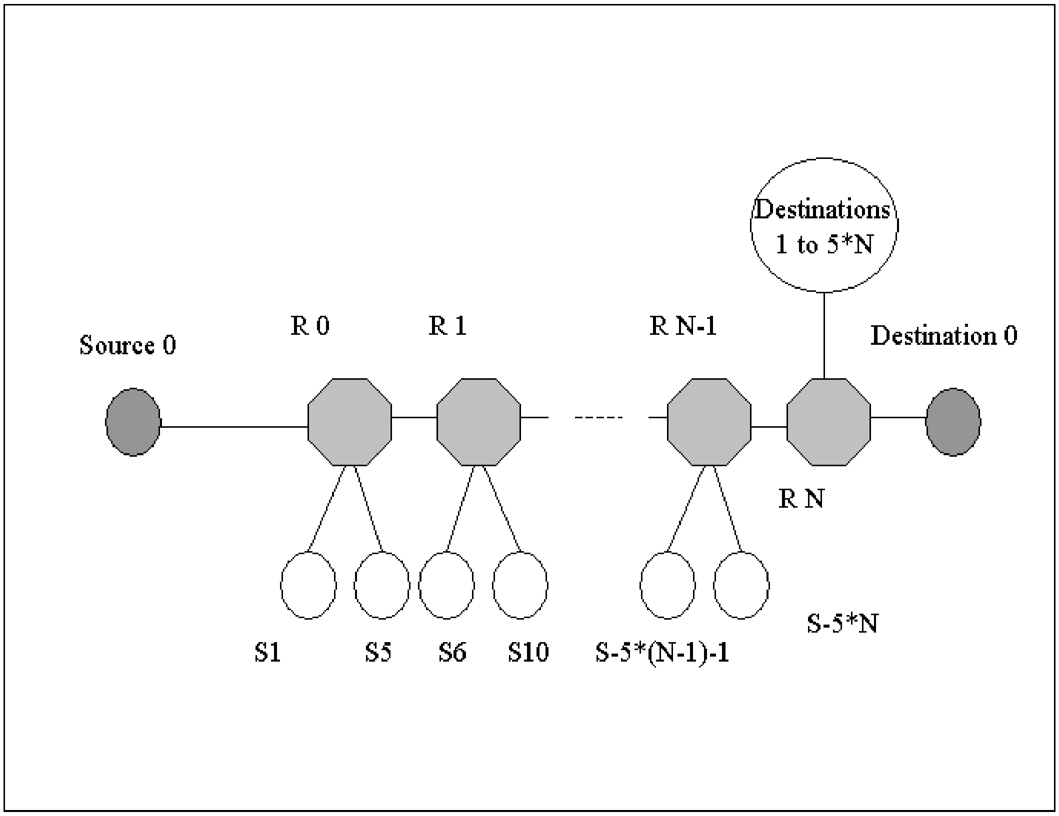

Figure A.5: Multiple Congested Links – Network Configuration

There are N+1 routers labeled R0 through RN. The links connecting the routers have a bandwidth of 10 Mbps and a link delay of 1 ms. Flows enter the network at routers R0 to R(N-1). All the flows leave the network at the router RN. At router R0, flow S0 enters the network. At router Ri flows S(i*5+1) to S((i+1)*5) enter the network. In each experiment set, the number of congested links was varied from 2 to 5.

Note that 0th flow has to contend for bandwidth with all other flows. The performance of PAB in multiple congested links is defined as the ratio of throughput of flow S0 to its SIR divided by the ratio of the sum of throughputs of all flows to sum of SIRs of all flows. This measure is the allocation ratio of flow S0:

Allocation Ratio AR (0) = [Throughput Ratio(0) / SIR Ratio(0)]

Four sets of experiments were done. In Experiment set 1, the flow S0 is a CBR UDP flow with its SIR set at 5 Mbps. The flows S1 to S(n*5) were also CBR UDP flows with SIR of 5Mbps. The flows were sending data at SIR. Figure A.6 shows the allocation ratio for the flow S0 vs. the number of congested links for equal fractions, AP fractions, GP fractions and simple RED. From the A.6 it is clear that the performance of simple RED is very poor. Among the three types of fractions, equal fractions performs worse than AP fractions or GP fractions. AP fractions and GP fractions are better suited for severe congestion.

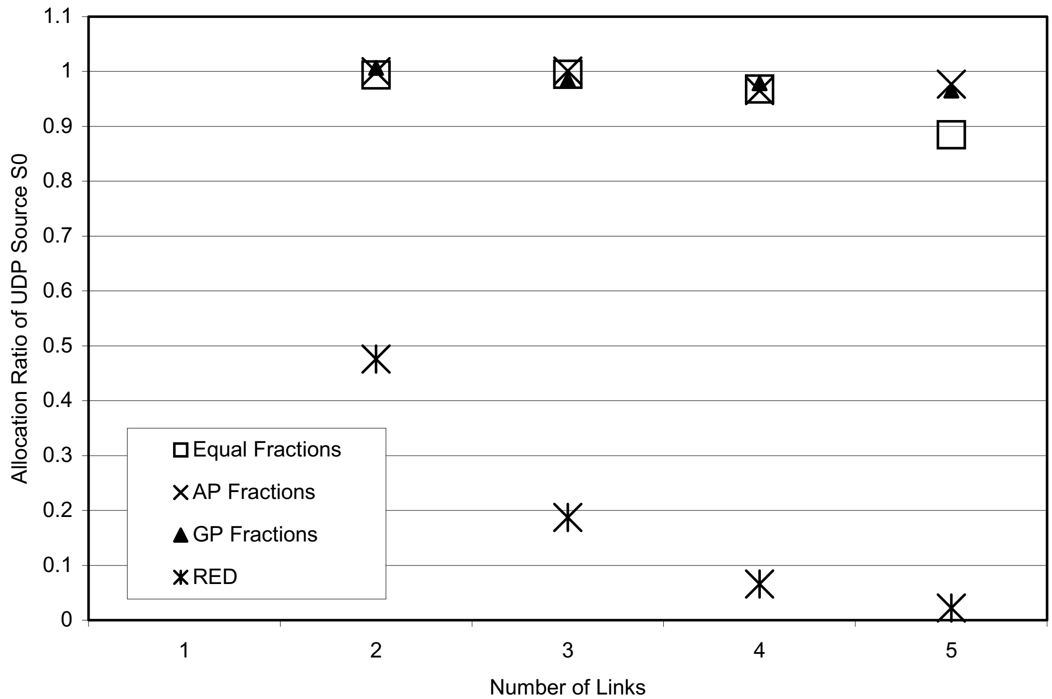

Figure A.6: Allocation Ratio of UDP Flow-0 in MCL with equal SIR

In the Experiment set 2, the flow S0 is a TCP Telnet application with its peak rate set at 5 Mbps. All the other flows S1 to S(n*5) were CBR UDP flows sending at their SIR of 5 Mbps. Figure A.7 shows the allocation ratio for the flow of Flow S0 vs. the number of links for equal fractions, AP fractions, GP fractions and simple RED.

The performance of simple RED with TCP as flow S0 is very poor and almost nil. Among the three types of fractions, performance variation occurs as the number of links increases. This is due to the fact that our technique is only an approximate implementation of PAB. Further TCP behavior varies widely depending on the threshold value and the actual fraction of the SIR currently allowed through the link.

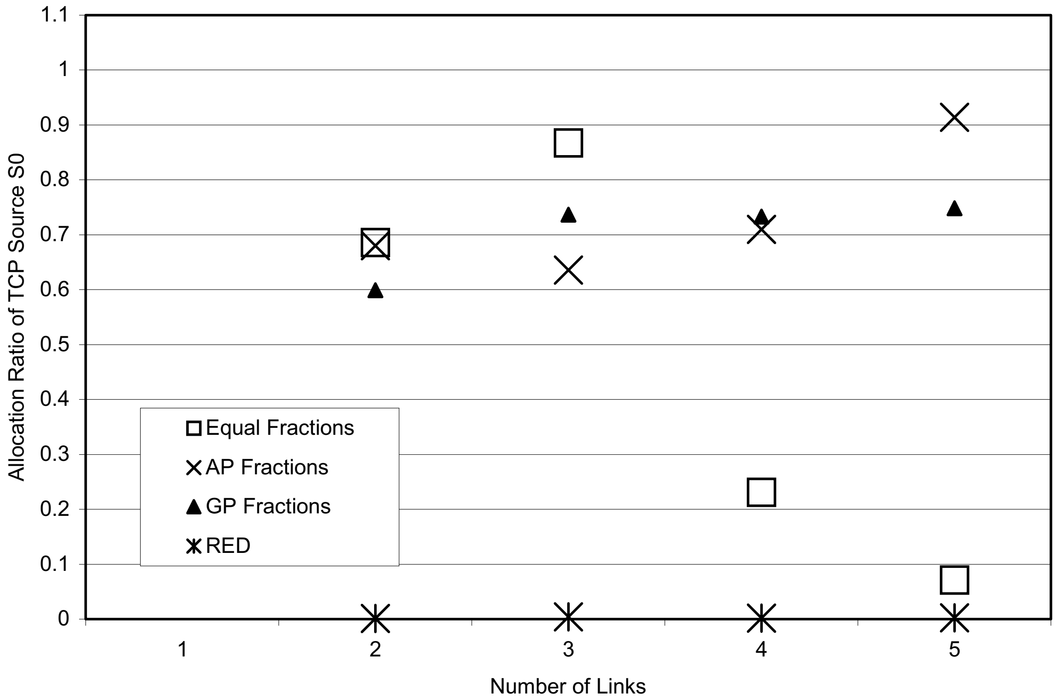

Figure A.7: Allocation Ratio of TCP Flow-0 in MCL with equal SIR

In the Experiment set 3, the flow S0 is a UDP flow with its SIR set at 5 Mbps. The flows S1 to S(n*5) were CBR UDP flows. The SIR of the flows were set to random values between 1 and 10 Mbps. Figure A.8 shows the allocation ratio for the flow of Flow S0 vs. the number of congested links for equal fractions, AP fractions, GP fractions and simple RED.

From Figure A.8, it is clear that the performance of simple RED is very poor. Among the three types of fractions, equal fractions performs worse than AP fractions or GP fractions. Again, during congestion AP fractions and GP fractions are better suited.

In the Experiment set 4, the flow S0 is a TCP Telnet flow with its peak rate set at 5 Mbps. All the other flows S1 to S(5*n) were the same as in Experiment set 3 with their SIRs varying randomly from 1 to 10 Mbps. Figure A.9 shows the allocation ratio for the flow S0 vs. the number of links for equal fractions, AP fractions, GP fractions and simple RED.

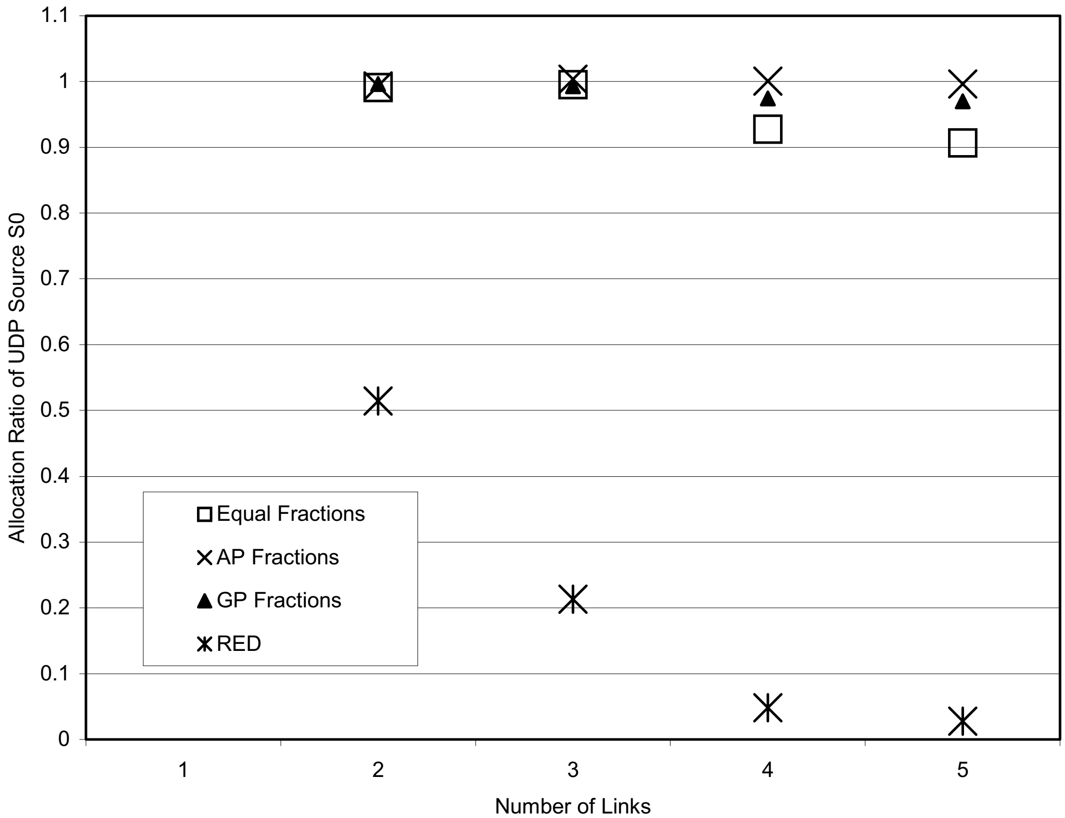

Figure A.8: Allocation Ratio of UDP Flow-0 in MCL with random SIR

Again, the performance of simple RED with TCP as flow S0 is very poor and almost nil. Among the three types of fractions, performance variation occurs as the number of links increases. Further TCP behavior varies widely depending on the threshold values and the actual fraction of the SIR currently allowed through the link.

The performance of equal fractions suffers significantly as the number of congested links increases. In equal fractions the highest priority DSCP is associated with a fraction value which is 1/8 th of SIR. So as severity of congestion increases, packets that are marked with any lower priorities are all dropped. Packets of the highest priority alone survive the congestion. Since now all packets are of the same priority, it becomes difficult to achieve PAB.

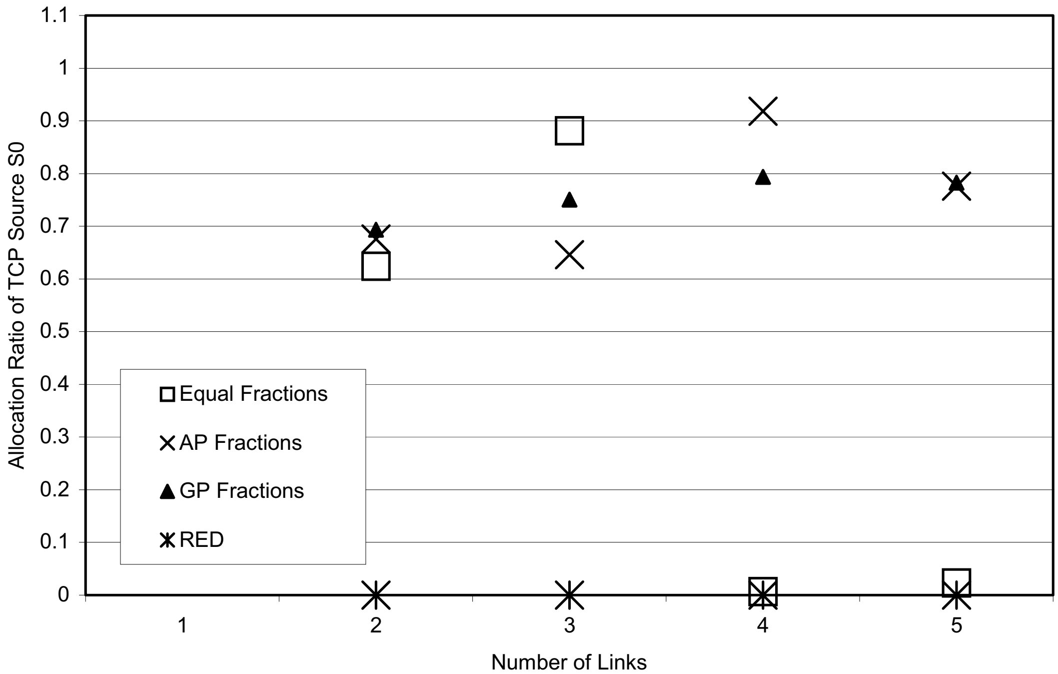

Figure A.9: Allocation Ratio of TCP Flow-0 in MCL with random SIR

The performance of equal, AP and GP fractions depends on the severity of the congestion in the network. When the congestion is less, all three schemes provide comparable performance. As congestion increases, the performance of equal fractions suffers. As the severity of the congestion increases AP fractions and GP fractions provide better performance.