A Selective Packet Discard Scheme for Supporting Internet QoS in Congested ATM Switches

Maurizio Casoni

Dept. of Engineering Sciences
University of Modena and Reggio Emilia
Via G.Campi 213/B
41100 Modena - I
casoni@dsi.unimo.it

D.E.I.S.
University of Bologna
Viale Risorgimento, 2
40136 Bologna - I
mcasoni@deis.unibo.it

Abstract

Previous papers have described and evaluated the performance of packet discard techniques for maintaining packet integrity during overload in ATM switches but without taking into account different classes of service.

In this paper a packet discard scheme is proposed with the goal to provide incoming flows of cells with output link bandwidth depending on the class of service they belong to. Also good performance in terms of throughput and fairness in bandwidth sharing among equal priority flows are considered as the requirements to meet.

In the scenarios here considered, the proposed discard scheme is shown to provide very good results and to represent a possible solution for dealing with the most demanding data flows during congestion intervals. Numerical results are reported and discussed for different operating conditions and traffic patterns.

1 Introduction

The switching in the Asynchronous Transfer Mode (ATM) [1], chosen by ITU-T as the basic transfer mode for the Broadband Integrated Services Digital Network (B-ISDN), is based on the fast packet switching technique. ATM networks are supposed to carry many different types of traffic with different bandwidth requirements. Applications range from continous ones transmitting informations at fixed rates to highly bursty which send large blocks of data at high peak rates but with long-term average rates relatively low. All applications are supposed to provide a set of traffic descriptors at the virtual circuit setup to describe the traffic sent with sufficient accuracy for routing and call acceptance. While many applications (audio, video, low speed data) are predictable enough for this strategy to work, many recent data applications involve interactive information retrieval with large data files (images) retrieved across the network on demand by a human user. The unpredictability of the human user and the high peak transmission rate required to provide acceptable interactive response imply that one must accept either a low network utilization or the possibility of overload periods during which the traffic sent to some network links exceeds their capacity. The fact that end-to-end protocols fragment their data in packets containing many ATM cells makes the impact of overload periods very dangerous, since a single lost cell can lead to the loss and retransmission of an entire transport level packet. Available bit rate (ABR) services are controlled by rate-based mechanisms but they work on round trip time intervals and so some other mechanisms must be implemented to cope with congestion occurring in short time intervals [2]. Packet discard schemes [3] [4] are ATM buffer management techniques to ensure high end-to-end throughputs [5] and fairness [6] [7] for bursty data applications during overload periods. Packet discard schemes are implemented in output buffered ATM switches and perform their task by selectively discarding cells from the information flow.

Other approaches to congestion control and fairness in ATM switches are based on the implementation of per-VC queueing and several algorithms and designs for the output controller have been proposed [8] [9]. Even if these proposals seem to be very effective some recent works have underlined possible weaknesses of traditional packet discard techniques, such as random early detection [10], when employed in real networks [11]. On the other hand other works, such as [12], confirm the effectiveness of early packet discard to improve the throughput. In conclusion, more efforts must be made to better understand the performance of these techniques and, in particular, to develop ever improved solutions.

TCP/IP networks do not guarantee any type of quality of service (QoS) since their delivery service, "best effort", implies that each IP data flow can have an indeterminate level of packet loss, delay, out-of-sequence. Since many multimedia applications have quite stringent requirements in terms of delay, loss and miminum bandwidth, a certain number of proposals have come out to provide some QoS to applications running over TCP/IP. According to [13] a possible classification of models for service differentiation is the following: relative priority marking, service marking, label switching [14], integrated services [15]/RSVP [16] and static per-hop classification.

In this paper a packet discard scheme is proposed to properly manage in congested ATM switches the flows of cells depending on the class of service their virtual circuit (VC) belongs to. IP packets are fragmented into cells and then transmitted over virtual circuits. Once a particular decision has been taken on the type of service class of a given flow of IP packets, the related virtual circuit must meet it at the best, also during congestion periods: thus, higher priority must be assigned to some VCs with respect to others for the output link bandwidth usage. To this end the packet discard scheme here proposed aims at providing two-class differentiation and, at same time, mantaining high throughput and fair exploitation of the bandwidth of the output link among equal-priority VCs.

The reported results, obtained from the simulation of a single switch output controller, show that these goals are met so that this scheme can be considered a possible solution to implement service differentiation in congested ATM switches.

The paper is organized as follows: section 2 describes the proposed discarding scheme, section 3 reports some numerical results for different scenarios and traffic patterns; finally conclusions are in section 4.

2 Selective Packet Discarding for Two Classes of Service

Some packet discard schemes have already been presented [3] [4] and evaluated [5] [6] [7] but none of them can discriminate among different classes of service. It is important to underline that we do not refer to a particular model of providing QoS but we just assume to divide incoming flows, associated to VCs, between low and high priority (two classes of service are supposed) and that the adopted model is the only responsible for this relationship.

The proposed scheme employs two thresholds, b_l (low) and b_h (high). The buffer controller tries to keep the occupancy level between b_l and b_h by activating or deactivating the VCs at packet boundaries according to their priority level. The goal of this packet discarding scheme is to provide higher priority VCs (i.e. cells, i.e packets) with most of the bandwidth, to optimize the bandwidth exploitation and to reach at the same time a good fairness in bandwidth sharing among equal priority VCs. VC de/activation depends on the number of successfully transmitted cells during a time interval, called the current window, which is compared with the number transmitted by the other VCs. Since the buffer occupancy level has a cyclic behaviour [5], two thresholds aim at improving the performance by increasing the number of packets successfully transmitted during each cycle. A virtual circuit is defined to be active if the buffer controller has decided to accept the cells of the current packet; otherwise, it is inactive.

Let n_i be the number of transmitted cells by VC i in the current window; A_m (I_M) be the smallest (largest) n_i from the set of currently active (inactive) VCs. The scheme here proposed is similar to the one presented in [6] with some variants to take priorities into account. Keeping in mind that during congestion all VCs send packets continually one beside the other, the scheme works as follows:

1. if the current buffer level exceeds b_h, at the packet boundaries of VC J make J inactive and always discard low priority cells;

2. if the current buffer level is below b_l, at the packet boundaries of VC J make J active;

3. if the current buffer level is between b_l and b_h, at the packet boundaries of VC J the buffer level is falling, J is inactive, $n_j < A_m$ and J is high priority make J active;

4. if the current buffer level is between b_l and b_h, the level is rising, J is active, $n_j > I_M$ and J is low priority make J inactive;

5. in all other cases do not change the state of J.

First two rules try to avoid that buffer overflows and underflows for optimizing the throughput and to save buffer space for high priority data. Third and fourth ones try to obtain fairness for equal priority VCs by turning on and off the VCs that have sent fewer or more cells in the current window, with respect to A_m and I_M. To handle fairness, the proposed scheme must keep the information regarding bandwidth usage of each VC in the current window. Also, it must maintain a record of the evolution of the buffer occupancy level in time in order to be able to decide if it is rising or falling.

Considering a homogeneous situation in which r VCs transmit data continually at a normalized rate of λ (that is, λ is the fraction of the link bandwidth required by a single virtual circuit), $r\lambda$ is defined as the overbooking ratio and for an overloaded link $r\lambda > 1$. It is assumed that all packets of all virtual circuits contain ℓ cells and let $k = \lfloor 1/\lambda \rfloor$ be the maximum number of virtual circuits that the link can handle without loss. Let B denote the number of cells the buffer can contain. The values of

B, b_l, b_h largely determine the goodput, defined as the fraction of the link's capacity used to carry complete packets, that it is possible to achieve on a link that has a buffer managed by this packet discard scheme.

If b_l and b_h are properly set to avoid overflow and underflow and the difference $(b_h - b_l)$ is large enough, it is reasonable to foresee a great difference in bandwidth usage between high and low priority data and a quite good degree of fairness among equal priority VCs. The reason is that in this condition the scheme can turn on and off the VCs depending on priorities and bandwidth usage without penalizing throughput.

3 Numerical results

A simulation program of the algorithm described previously has been developed to study the performance in terms of per VC bandwidth usage, output link goodput and fairness in bandwidth usage. The bandwidth usage is the figure of merit for the level of quality assigned to each VC in presence of congestion and it must respect the class of service (priority) of the related user data flow. The discarding scheme will then be mainly tested on this parameter. Besides, it is worth reminding that the proposed scheme has two thresholds which aim at increasing the number of succesfully transmitted packet per cycle, thus improving the goodput.

The simulation requires to set up several parameters: the simulation duration is 1000 packet times long, the current window is 200 packet times. All numerical results assume $\ell = 171$ to represents one ATM adaptation layer (AAL) 5 packet.

Assuming a 600 Mbit/s output link, the bandwidth usage is considered in the overbooking ratio range up to 3 Gbit/s which is a likely scenario, which means $r\lambda$ in the range [1..5]. The scheme has been evaluated in different operating scenarios and traffic patterns, namely, with homogeneous sources sending data at the same rate, or normalized rate λ, and with non homogeneous sources which have been grouped in two different sets characterized by different data rates, λ_1 and λ_2 respectively.

overbooking ratio	2	3	4	5
Case 1	0.9772	0.9825	0.979	0.9786
Case 2	0.9977	0.9993	0.9995	0.998

Table 1: Goodput as a function of the overbooking for two analyzed cases.

Next paragraphs will report some numerical results for these two main scenarios.

3.1 Homogenous sources

Two situations are here analyzed. Both refer to an homogeneous situation meaning that all VCs have the same rate. Case 1 has $k = 12$, $B = 512$, $b_h = 256$, $b_l = 64$ and 1/3 of the VCs are high priority. Case 2 corresponds to $k = 10$, $B = 812$, $b_h = 512$, $b_l = 128$ and 1/4 of the VCs are high priority. In all cases high priority VCs have been randomly chosen.

Figures 1 and 2 refer to case 1 and show the bandwidth usage for each VC for an overbooking ratio equal to 2 and 3 respectively. In Figure 1 high priority VCs belong to the number set $(3, 6, 9, 11, 15, 18, 20, 23)$: most of the bandwidth is assigned to higher priority VCs and within the two classes of services a significant level of fairness is obtained. In Figure 2 high priority VCs belong to the number set $(2, 5, 8, 11, 15, 17, 20, 23, 27, 30, 32, 35)$ and previous considerations still hold.

Figures 3 and 4 refer to case 2 and show the bandwidth usage for each VC for an overbooking ratio equal to 3 and 5 respectively. In Figure 3 high priority VCs belong to the number set $(2, 5, 8, 11, 15, 18, 22, 26)$ while in figure 4 belong to $(3, 6, 9, 13, 17, 21, 26, 29, 34, 37, 40, 46, 49)$ and again the bandwidth is assigned mostly to high priority VCs with a reasonable fairness.

Table 1 shows the goodput values reported for the two considered cases and underlines the effectiveness of the proposed scheme in optimizing the output link.

The results obtained mainly match the proposed target of assigning the output link to higher priority VCs first, keeping the goodput as high as possible and trying to provide an equal sharing of the link among equal priority VCs.

3.2 Non homogeneous sources

In this section the packet discard scheme is studied in a different scenario. Let us assume that VCs have two different rates, λ_1 and λ_2 with $\lambda_1 > \lambda_2$. It is also assumed that VCs still transmit packets of the same length, $\ell = 171$. As a consequence, the packet transmission time is different for VCs with different rates and $\ell/\lambda_1 < \ell/\lambda_2$. Therefore, during the transmission time $T_2 = \ell/\lambda_2$ more than one packet boundaries per each VC can occur for faster VCs. This means that faster VCs ask the buffer controller for more accesses to the buffer than slower ones and this may lead to some unfairness between fast and slow VCs.

In the following cases, high priority is assigned to faster VCs, assuming that they are used for time critical applications or because users simply pay for that. Of course, other assumptions and mixes can be made but they will not be discussed here.

In Figure 5 we assume $\lambda_1 = 0.5$, $\lambda_2 = 0.2$, $r_1 = 5$ and $r_2 = 5$ giving an overbooking ratio equal to $r_1\lambda_1 + r_2\lambda_2 = 3.5$. Basically, all the bandwidth is taken by the r_1 sources and a quite good level of fairness is achieved among them. On the other hand, the r_2 sources get nothing but, actually, this is just what we want: priorities have been set up to assign resources, i.e. bandwidth, to specific users, in particular during the short congestion intervals.

In Figure 6 $\lambda_1 = 0.5$, $\lambda_2 = 0.2$, $r_1 = 7$ and $r_2 = 5$ giving an overbooking ratio equal to $r_1\lambda_1 + r_2\lambda_2 = 4.5$. Here the set $(1, 3, 5, 7, 8, 11, 12)$ represents high priority VCs and again they take all the bandwidth and they quite fairly share it. Table 2 shows the values of goodput each VC gets for this last case: the total 99% is composed of contributions provided by high priority VCs while the low priority ones do not send succesfully any packet. Since the total goodput is so high, these values of goodput for each VC roughly correspond to the bandwidth usage reported in figure 6. This means that the bandwidth assignment based upon this scheme is performed in such a way that basically the whole bandwidth is used for transmitting complete packets.

VC number	1	2	3	4	5	6	7	8	9	10	11	12	Total
Goodput	0.16	0.0	0.13	0.0	0.13	0.0	0.16	0.12	0.0	0.0	0.16	0.1462	0.99

Table 2: Goodput per VC for the case of fig. 6.

A different case is reported in figure 7 where $\lambda_1 = 0.2$, $\lambda_2 = 0.1$, $r_1 = 5$ and $r_2 = 30$ and the overbooking ratio is $r_1\lambda_1 + r_2\lambda_2 = 4$. Here there is a greater number of VCs, 35, than two previous cases and also the high priority VCs $(2, 13, 23, 28, 33)$ are 1/7 of the total. Again, as desired, they get most of the bandwidth but here low priority VCs can take small portions of it.

4 Conclusions

In this paper a selective packet discard scheme has been proposed for managing congestion in output buffered ATM switches taking into account the class of service of the incoming data flows. It gives output bandwidth to higher priority VCs first and here two-level priority data streams have been considered.

The analyzed scheme tries to keep the buffer occupancy level between two thresholds as long as possible by de/activating the VCs depending on the number of successfully transmitted packets in a time window and on their priority, in order to provide high goodput and good fairness in bandwidth sharing among equal priority VCs. In other words, the goal is not to avoid congestion, in fact congestion already occurs being $r\lambda > 1$, but at least to avoid overflow, i.e cell losses, and underflow, i.e. output link inactivities, providing a reasonable fairness among equal priority VCs.

For the scenarios here taken into account the proposed scheme has been shown to meet the mentioned target and, in particular, to be capable of managing successfully different priority data flows without penalizing throughput or fairness.

The above results have been obtained through simulation so that it is not straightforward to foresee the exact behaviour of the scheme in a real network. As a matter of fact, realistic scenarios can be more complex than the ones here presented and analyzed. Combinations of different packet lengths with many different data rates and priorities are possible and this is current work.

The subject of packet discarding schemes is still open to discussion about their effectiveness, as recent literature has underlined, so that many more investigations and possibly test-beds are necessary to better clarify this issue. However, these first results show that is possible to set up a packet discard scheme whose task is to provide class-dependent bandwidth assignment and high throughput with an acceptable fairness.

References

[1] M. De Prycker, "Asynchronous Transfer Mode: a Solution for B-ISDN", *Ellis-Horwood*, II edition, 1993.

[2] R. Jain, "Congestion Control and Traffic Management in ATM Networks: Recent Advances and a Survey", *Computer Networks and ISDN Systems*, vol. 28, no. 13, pp. 1723-1738, October 1996.

[3] S. Floyd, R. Allyn, "Dynamics of TCP Traffic over ATM Networks", *Computer Communication Review*, vol. 24, no. 4, 10/94.

[4] J.S. Turner, "Maintaining high throughput during overload in ATM switches", Proc. of Infocom '96, 24-28 March 1996, San Francisco, CA (USA).

[5] M. Casoni, J.S. Turner, "On the Performance of Early Packet Discard", *IEEE Journal on Selected Area in Communications*, special issue on *Advances in ATM Switching Systems for B-ISDN*, vol. 15, no. 5, pp. 892-902, June 1997.

[6] M. Casoni, "Fair Packet Discard as Improvement of Early Packet Discard", Proc. of Globecom'98, 8-12 November 1998, Sydney (AUS).

[7] S. Chan, E.W.M. Wong, K.T. Ko, "Fair Packet Discarding for Controlling ABR Traffic in ATM Networks", *IEEE Transactions on Communications*, vol. 45, no. 8, pp. 913-916, August 1997.

[8] L. Benmohamed, Y.T. Wang, “A Control-Theoretic ABR Explicit Rate Algorithm for ATM Switches with Per-VC Queueing”, Proc. of Infocom’98, 29 March - 2 April 1998, San Francisco, CA (USA).

[9] Y. Chen, J.S. Turner, “Design of a Weighted Fair Queueing Cell Scheduler for ATM Networks”, Proc. of Globecom’98, 8-12 November 1998, Sydney (AUS).

[10] S. Floyd, V. Jacobson, “Random Early Detection Gateways for Congestion Avoidance”, *IEEE/ACM Transaction on Networking*, pp.397-413, 1993.

[11] M. May, J. Bolot, C. Diot, B. Lyles, “Reasons not to deploy RED”, Proc. of 7th International Workshop on Quality of Service, 31 May-4 June 1999, London (U.K.).

[12] R. Goyal, R. Jain, M. Goyal, S. Fahmy, B. Vandalore, “Traffic Management for TCP/IP over Satellite ATM Networks”, *IEEE Communications Magazine*, pp.56-61, March 1999.

[13] S. Blake, D. Blake, M. Carlson, E. Davies, Z. Wang, W.Weiss, “An Architecture for Differentiated Services”, *Request for Comments 2475*, ftp://ftp.isi.edu/in-notes/rfc2475, December 1998.

[14] R. Callon, G. Swallow, N. Feldman, A. Viswanathan, P. Doolan, A. Fredette, “A Framework for Multiprotocol Label Switching”, Internet Draft, draft-ietf-mpls-framework-02.txt, August 1997.

[15] B. Braden, D. Clark, S. Shenker, “Integrated Services in the Internet Architecture: an overview”, *Request for Comments 1633*, June 1994.

[16] R. Braden, L. Zhang, S. Berson, S. Herzog, S. Jamin, “Resource ReSerVation Protocol (RSVP) - version 1 functional specification”, *Request for Comments 2205*, September 1997.

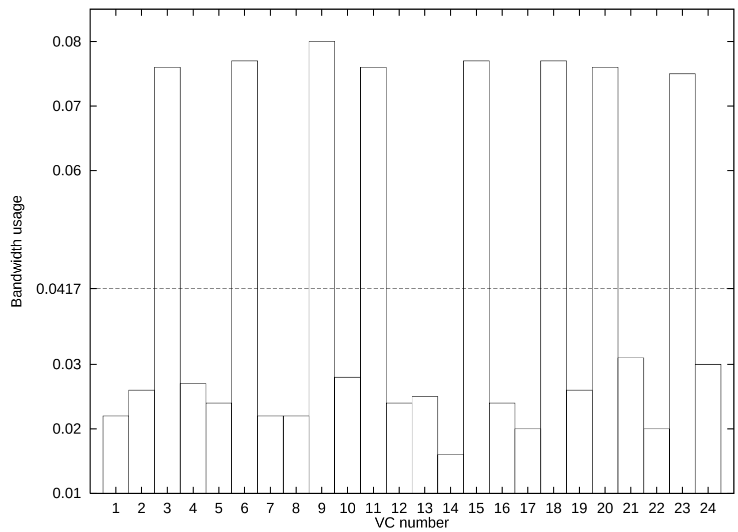

Figure 1: Bandwidth usage for each virtual circuit; $k = 12$, $r\lambda = 2$, $B = 512$, $b_h = 256$, $b_l = 64$; dotted line represents the ideal sharing corresponding to 1/24 of the output link.

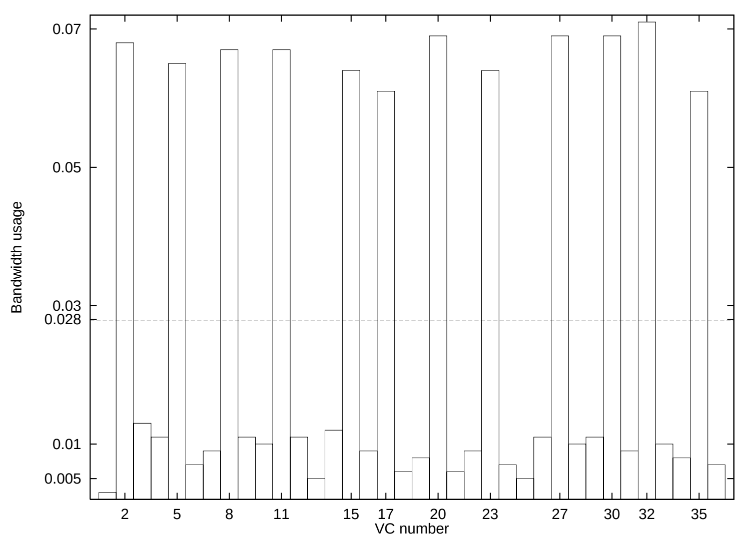

Figure 2: Bandwidth usage for each virtual circuit; $k = 12$, $r\lambda = 3$, $B = 512$, $b_h = 256$, $b_l = 64$; dotted line represents the ideal sharing corresponding to 1/36 of the output link.

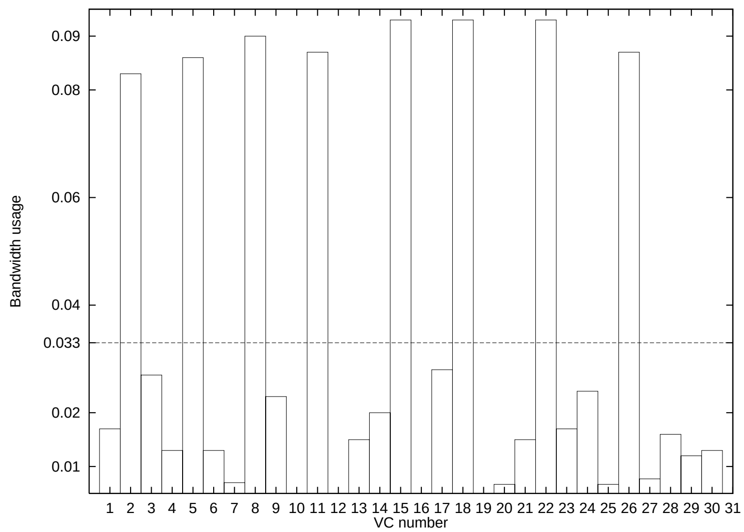

Figure 3: Bandwidth usage for each virtual circuit; $k = 10$, $r\lambda = 3$, $B = 812$, $b_h = 512$, $b_l = 128$; dotted line represents the ideal sharing corresponding to $1/30$ of the output link.

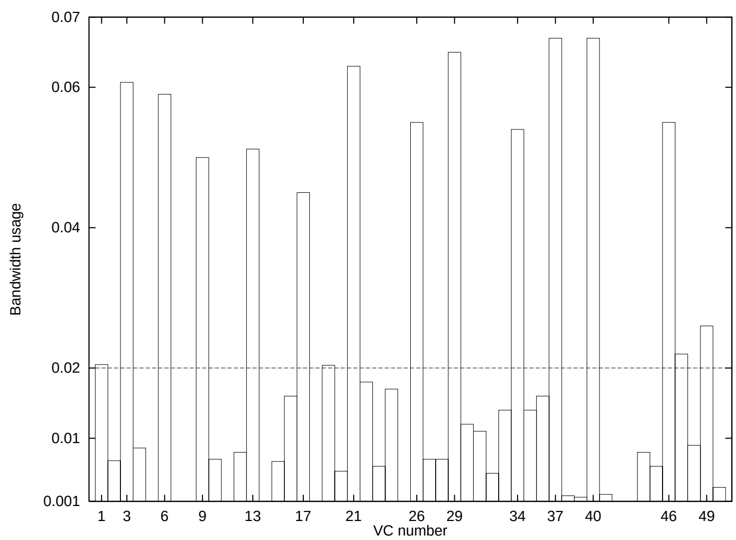

Figure 4: Bandwidth usage for each virtual circuit; $k = 10$, $r\lambda = 5$, $B = 812$, $b_h = 512$, $b_l = 128$; dotted line represents the ideal sharing corresponding to $1/50$ of the output link.

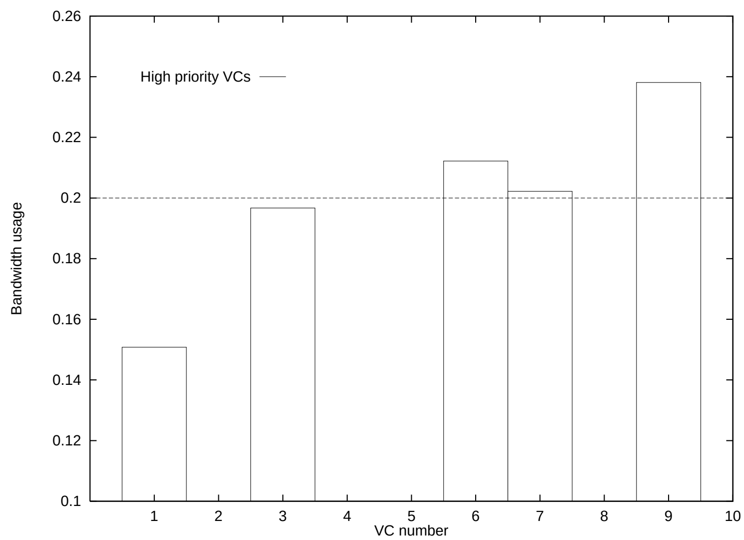

Figure 5: Bandwidth usage for each virtual circuit; $k_1 = 2$, $k_2 = 5$, $r_1\lambda_1 + r_2\lambda_2 = 3.5$, $B = 512$, $b_h = 256$, $b_l = 64$; dotted line represents the ideal sharing among high priority VCs (1,3,6,7,9) corresponding to 1/5 of the output link.

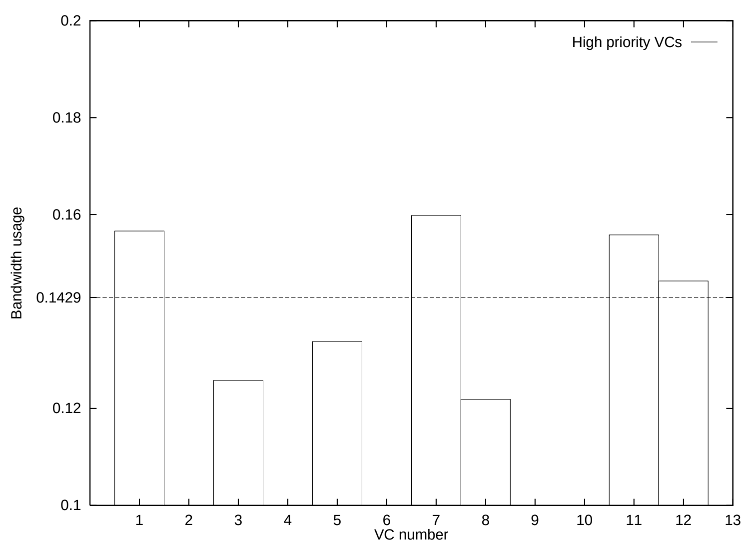

Figure 6: Bandwidth usage for each virtual circuit; $k_1 = 2$, $k_2 = 5$, $r_1\lambda_1 + r_2\lambda_2 = 4.5$, $B = 512$, $b_h = 256$, $b_l = 64$; dotted line represents the ideal sharing among high priority VCs (1,3,5,7,8,11,12) corresponding to 1/7 of the output link.

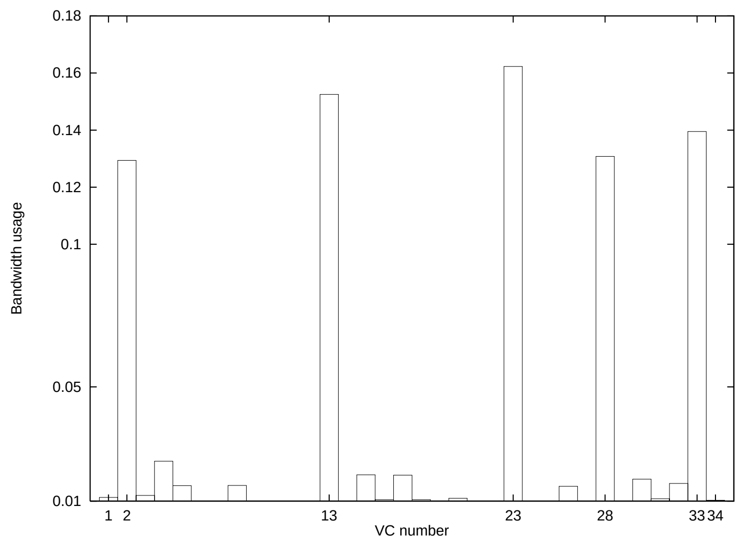

Figure 7: Bandwidth usage for each virtual circuit; $k_1 = 5$, $k_2 = 10$, $r_1\lambda_1 + r_2\lambda_2 = 4$, $B = 512$, $b_h = 256$, $b_l = 64$; (2,13,23,28,33) are the high priority VCs.