The Effect of Traffic Models on the Performance of Space Priority Queueing in ATM Switches

Karen L. Reid and Richard B. Bunt
Department of Computational Science
University of Saskatchewan
Saskatoon, Canada S7N 0W0

September 15, 1993

Abstract

High-speed packet switched networks such as B-ISDN require techniques to manage the flow of traffic into and through the network. Each switching node and network access node is susceptible to overload. End-to-end congestion control will do much to ensure that overload periods are minimized, but intelligent buffering decisions at internal nodes may aid in managing congestion by shifting cell loss to lower priority traffic.

This paper investigates the effect of traffic characteristics on the performance of a typical space priority queueing algorithm. Simulation was used to examine four traffic models and to extend previous work with analytic models using Poisson traffic. The results of the simulations with bursty (more realistic) traffic show that the Poisson model leads to drastically optimistic cell loss. Other results indicate that cell loss level is most strongly affected by the burstiness of the traffic and by the load on the switch. Nevertheless, even under bursty traffic, space priority algorithms, when combined with other congestion control methods, can aid in supporting acceptable quality of service for the users of the network.

Keywords: B-ISDN, ATM, buffer management, traffic modeling, performance evaluation, simulation

1 Introduction

The high-speed integrated service networks of the future will support a wide variety of applications. The characteristics of the network traffic generated by these applications will be equally diverse, and this diversity presents new challenges. Several characteristics of the traffic heavily affect network performance and should thus influence decisions about the types of control measures the network can use. Reserving channels for high priority connections mitigates some congestion problems but leads to inefficient use of bandwidth. Traffic is expected to be bursty, a feature that aggravates congestion problems. The impact of traffic characteristics on network performance stresses the importance of realistic models in studies of network congestion.

The characteristics that determine the shape of the traffic are difficult to quantify. Burstiness is sometimes defined as $B = peak\ bit\ rate/mean\ bit\ rate$. This value does not take into account the length of the bursts or the proportion of the traffic that occurs during burst periods. It also ignores correlation factors in the traffic. In this paper, the term *burstiness* will refer to the measure B, and the term *shape* will include burstiness and the other traffic characteristics described above.

A single 8x8 output-buffered Asynchronous Transfer Mode (ATM) switch is simulated for this study. The input and output links carry 53 byte ATM cells and have a total bit rate of 150 Mbps. This data rate corresponds to proposals for Broadband ISDN networks [1]. Space priority queueing [8, 9] is performed at the output buffers using the *push out*

algorithm. The traffic stream from all of the 8 inputs consists of cells from one of four distinct traffic models destined for a single output. Modeling only one output port yields all the interesting information when separate buffers are attached to each output since the priority queueing algorithms operate at the point where cells enter the output buffer.

The goal of this research was to examine several traffic models that exhibit different traffic characteristics and to study their effect on the performance of a representative space priority queueing algorithm. The rest of the paper is organized as follows: Section 2 describes priority queueing and the push out algorithm. The various traffic models are introduced in Section 3. The simulation experiments and the results obtained are discussed in Section 4, and Section 5 presents the conclusions of the work.

2 Space Priority Queueing

Managing congestion is a significant component of meeting quality of service (QOS) demands. Congestion can and will be managed at several levels in future networks. The most prominent approach, end-to-end congestion control, adjusts the rate at which sources transmit to prevent or react to overloading on the network. Access control algorithms and end-to-end congestion control algorithms may not be sufficient to prevent congestion in a high speed integrated services network, and more intelligent decisions about dropping cells inside the network can help to support end-to-end algorithms. For this paper, congestion is addressed inside the network through the use of priority-managed buffers at the output

links of internal switches.

When the capacity of an output link is exceeded by a short-term burst of cells, a buffer (or queue) can be used to reduce the cell loss. Choosing which cells to drop in the event of buffer overflow is the responsibility of the buffer management algorithm. If QOS requirements are to be met, the choice should be based on the priority of the cells in the buffer. (The CCITT has reserved one bit in the ATM header for cell loss priority [1].) Cell loss priority or *space priority* algorithms address the problem of which cells to drop as the buffer fills. The *push out* algorithm operates only when the buffer is full. When a low priority cell arrives at a full buffer, it is dropped. When a high priority cell arrives at a full buffer and a lower priority cell is in the buffer, then the lower priority cell is "pushed out," and the arriving high priority cell is added to the buffer. This approach is *work conserving* [9]; it drops cells only when the buffer is full. Therefore, the total cell loss is the same as if no space priority queueing scheme were used.

The push out algorithm has been studied analytically using Poisson traffic models and two priority classes [5, 8]. Lin and Sylvester [9] also take into account possible correlation between the two traffic classes. All three studies show that when two priority classes contribute equally to the load, the cell loss for the higher priority traffic (Class 1) is dramatically reduced while the cell loss for the lower priority traffic (Class 2) increases slightly.

The simulation model used for this study gave similar results for the push out algorithm

under a Poisson traffic model, but when bursty traffic models were used cell loss was much higher. Saito [12] and García and Casals [3] studied the push out algorithm under an MMPP traffic model. Both studies show results similar to [5] and to results using MMPP with the simulation model in this study.

A previous study [11] investigated the performance of two other priority queueing algorithms: partial buffer sharing, and selective packet discarding. Traffic streams composed of a mixture of the models described in the next section were used to compare the performance of the three algorithms.

3 Traffic Models

Traffic on high-speed networks will be bursty and correlated [4, 7]. Simple traffic models such as Poisson do not capture important characteristics of the traffic. Furthermore, since the mandate of high-speed networks such as B-ISDN is to integrate a wide variety of services, the traffic characteristics of each class of services will differ. Results of studies based on simplistic assumptions (such as Poisson traffic models) may have little relevance to practical networks. An important goal of this research was to incorporate a wider range of traffic types into the investigation.

Three traffic models were used, representing three broad classes of traffic: voice, video, and data (each model represents an aggregate stream of traffic). Much work has been done in characterizing traffic of each of these three types in different networks and the

three traffic models used here were drawn from previous work. The Poisson model was also simulated, to provide performance comparisons to the other three models.

MMPP (Markov Modulated Poisson Process) has been used to characterize telephone voice traffic [6]. This model captures burstiness still present in multiplexing of many voice sources. In general, MMPP is an n-state model where the holding times and interarrival times in each state are exponentially distributed. A two-state model was used in this study.

The model used for data traffic is similar to MMPP in that it is also a two-state model, but it has a deterministic rate for cell transmission during the active state and does not transmit cells during the idle state. The design of the data model is motivated by the recognition that bursts of data traffic usually consist of a single transfer of a larger piece of information and that cells generated by this transfer will likely be transmitted at a constant rate [13, 2]. The model chosen has an exponentially distributed idle period and the number of cells in a burst is geometrically distributed. During a burst, cells are transmitted at the peak rate.

The characteristics of video traffic depend to a large extent on the encoding algorithm used. The video model used in this simulation study is one developed by Maglaris *et al.* [10] based on measurements. The bit rate of a single source during the nth frame is modeled by a first-order autoregressive Markov process. The length of a burst is based on the number of frames per second and the cell rate during a burst is based on the number of cells (or bits) per frame. By varying these two values, traffic streams of different load levels can

be obtained. This does not violate the characteristics of the model since Maglaris *et al.* [10] note that their model assumptions hold for a wide variety of video sources. Although the burst length of a video source is expected to be fairly long (on the order of 40 ms), this length is infeasible for practical simulation runs. A burst length of 1.7 ms was used for this study. Longer burst lengths lead to smoother traffic at the switches, and if call acceptance algorithms ensure that the total traffic presented to the network does not exceed the network's capacity, then this should lead to lower cell loss.

The four models were modified to enable them to label cells as high or low priority. Because they all model aggregate traffic, low priority cells are randomly distributed throughout the traffic stream, and the proportion of low priority cells to high priority cells is determined by a parameter to the model.

Figure 1 highlights the different characteristics of each of the traffic models at a load level of 0.1. The mean bit rate for each aggregate traffic stream is 15 Mbps (40 cells per 400 slots) but the peak bit rates vary. The variation in arrival rates is much greater for the data model than for the other two models. Of the three bursty models, the video model shows the least variation in the traffic rate, and this variation occurs over longer periods of time than in the other two models. The dotted lines in each graph indicate the mean bit rate of the traffic model.

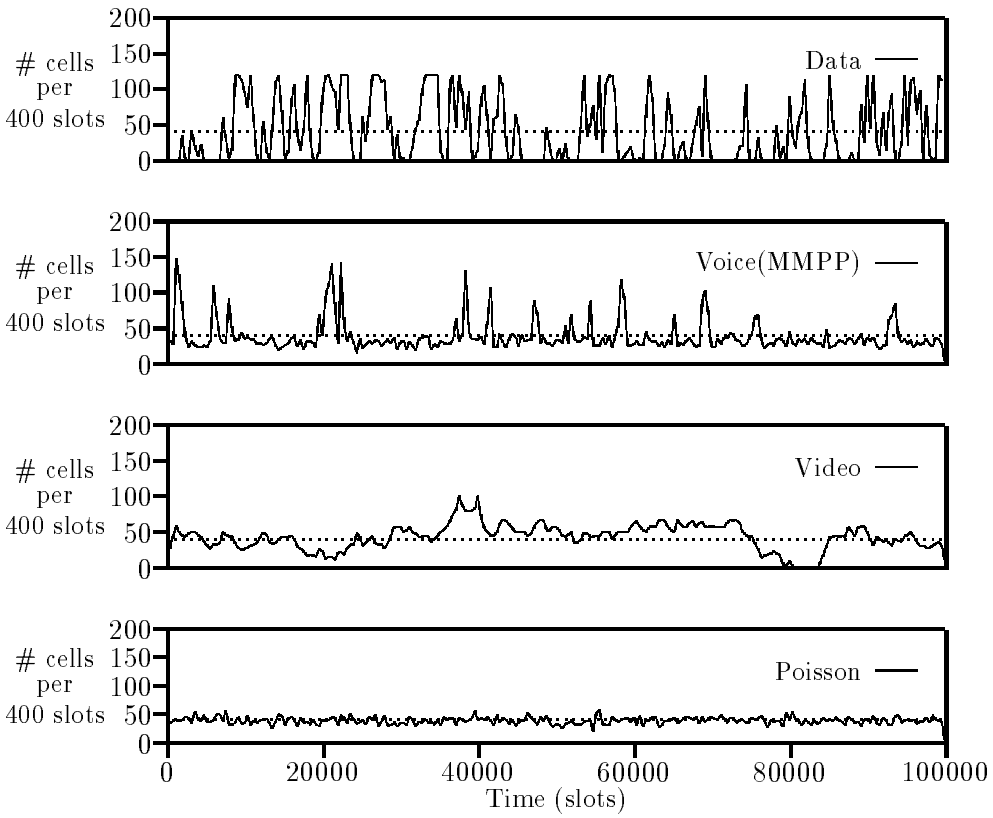

Figure 1: Shape of Four Traffic Models (load level = 0.1)

4 Simulation and Results

The variables of the simulations include the buffer size at the ATM switch, the load level, and the traffic model. The parameters of the traffic models are varied to model different levels of burstiness, burst length, and proportion of traffic in each priority class. Space priority queueing (with the push out algorithm) is applied at the switch buffer and performance is assessed in terms of cell loss. In most cases, confidence intervals were small ($< \pm 5\%$) with respect to the differences between results. Error bars are included in the

figures where the confidence intervals are larger. Where the confidence interval includes zero, the lower part of the error bar is indicated with ellipsis.

4.1 Total Cell Loss

The total cell loss experienced at a switch buffer depends on the size of the buffer, the load on the switch, and the shape of the traffic. The following figures illustrate the effect of each of these factors on the performance of the queueing algorithm.

Figure 2 compares the cell loss of the four traffic models at a load of 0.8. The Poisson model generates much smoother traffic than the other three models, so it needs only a small buffer (25-50 cells) to ensure low cell loss. It was not possible to simulate the Poisson model for buffer sizes larger than 40 because the rare event of dropping a cell ($< 10^{-10}$) requires prohibitive simulation times. When the traffic is bursty, and the burst length is longer than the buffer size, much higher cell loss occurs. In this figure, the burst length for the data model and the voice model (MMPP) is 357.1 slots (1 ms); the video model has a burst length of 607 slots (1.7 ms).

The data model shows the highest cell loss because all arrivals occur during the burst state. The buffer will fill, overflow, and then empty after the burst is finished. The voice model with the same burstiness shows less cell loss than the data model because a significant number of cells arrive during the non-burst state when the buffer does not fill. This may explain why the rate of decreasing cell loss is slightly greater for voice than

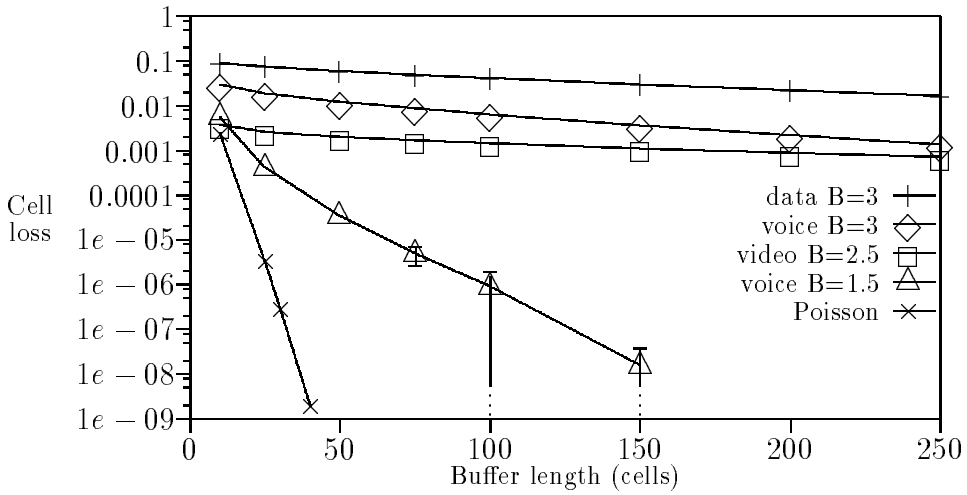

Figure 2: Total Cell Loss (load = 0.8)

for the other bursty models: the increase in buffer size prevents dropping cells during the non-burst state or during shorter bursts.

Note that when the burstiness is 1.5 for the voice model, the cell loss is still several orders of magnitude greater than that for Poisson traffic. The effect of reducing the burstiness of the voice model is the reduction of the magnitude of the peak rate. It does not alter the proportion of traffic that arrives during a burst state. This means that although the peak rate is lower, and the length of the bursts remains the same, the bursts occur more frequently.

Figures 3 and 4 show the impact of burstiness and the importance of quantifying it accurately. For each curve, the traffic source has the same fixed peak bit rate (Bp) and

fixed mean bit rate (Bm), and the load is the same in both graphs. The only difference between the graphs is the degree of burstiness. The buffer is fixed at a size of 50 cells.

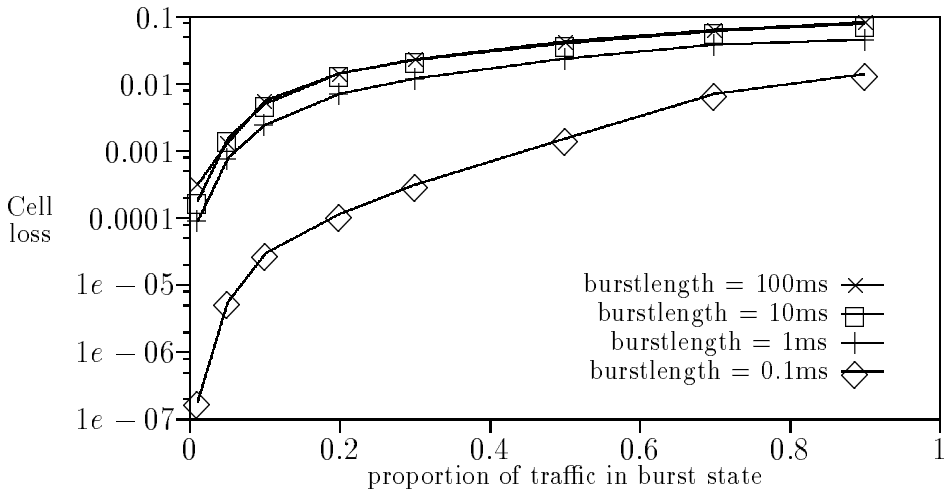

Figure 3: Total cell loss for varying burst lengths in Voice (MMPP) model (load = 0.8, B = 3)

When the burst length exceeds the buffer size (0.1 ms = 36 slots), the cell loss is several orders of magnitude greater than it is when the burst length is longer than the buffer. There is a small increase in cell loss from a burst length of 1 ms to a burst length of 10 ms (the confidence intervals do not overlap). As the burst length grows, however, a saturation point is reached where the cell loss no longer increases. There is no statistical difference between the two curves with the longest burst lengths. Also, there is little difference between the results for $B = 3$ and $B = 4$ for the larger burst lengths. This suggests that burstiness is an important factor, but when the burst length is much larger than the buffer size, the cell

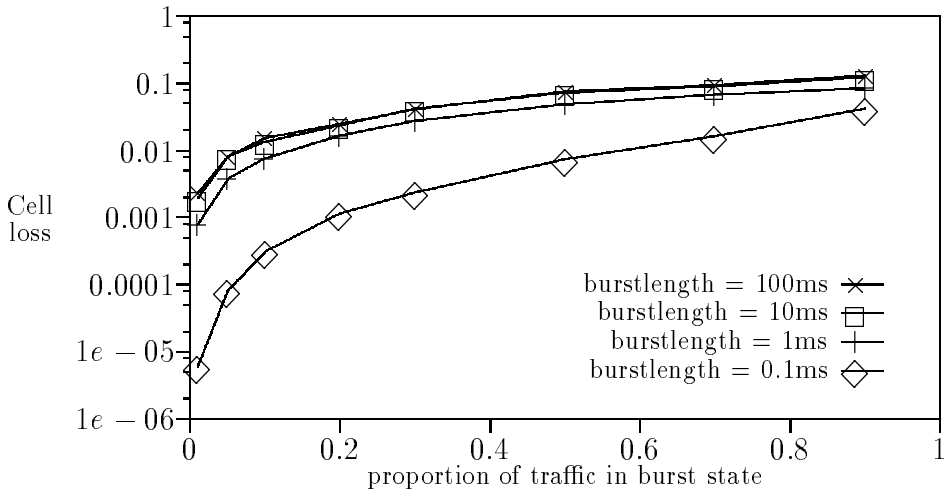

Figure 4: Total cell loss for varying burst lengths in Voice (MMPP) model (load = 0.8, B = 4)

loss is largely insensitive to burstiness and burst length.

4.2 Distribution of Cell Loss

Figures 5, 6, and 7 demonstrate that despite the higher total cell loss experienced by the bursty traffic models, the push out algorithm can still significantly protect the high priority traffic (Class 1) from excessive cell loss while increasing the cell loss of Class 2 traffic by less than an order of magnitude. (However, an increase in cell loss from 1% to 5% may still be considered a significant performance penalty.)

It is clear from these graphs, as well as from other studies using Poisson traffic, that when the traffic is smoother, not only is the total cell loss lower but the push out algorithm

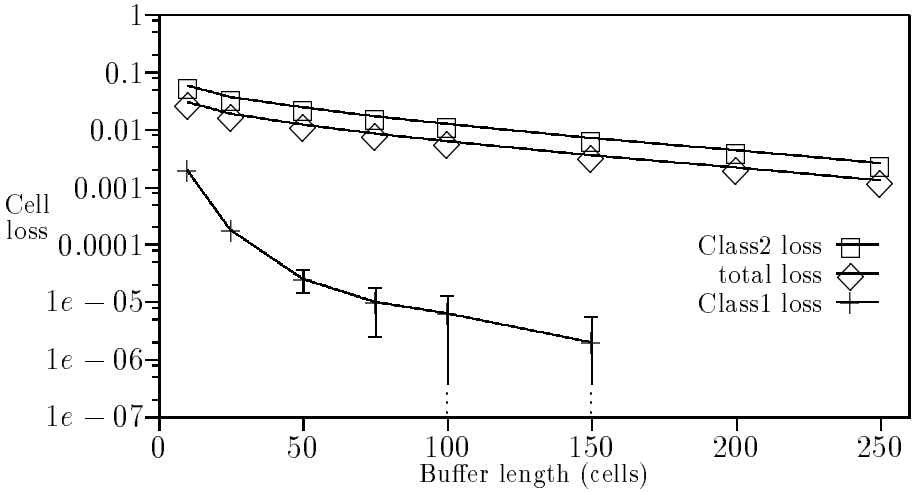

Figure 5: Voice (MMPP) cell loss by class (load = 0.8)

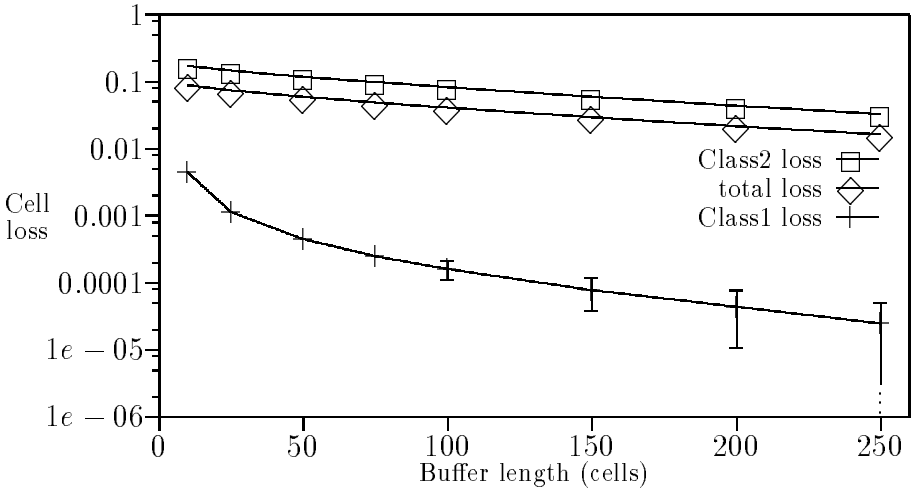

Figure 6: Data cell loss by class (load = 0.8)

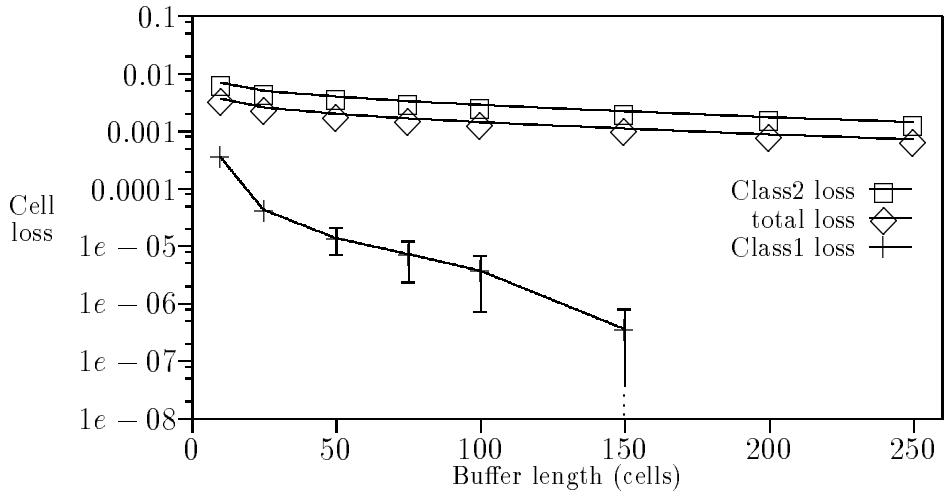

Figure 7: Video cell loss by class (load = 0.8)

is also much better able to protect high priority traffic. Note the difference in Class 1 cell loss in each of the figures. This will, in general, make a higher admissible load possible without adversely affecting the QOS.

The cell loss for the high priority traffic in the data model does not decrease as rapidly as with the other models when the buffer size increases. While this is caused partly by higher total cell loss, the shape of the traffic plays a larger role. When a burst begins, the buffer fills, and for the duration of the burst, few additional cells will find space in the buffer. Therefore, the high priority cells quickly replace all the low priority cells in the buffer. Until the burst ends, few low priority cells will gain access to the buffer. This means that many high priority cells will be lost.

Figures 8, 9, and 10 show that the proportion of high priority cells strongly affects the performance of the priority queueing algorithm. The total cell loss remains the same but when the traffic is made up of more low priority cells, a high priority cell arriving at a full buffer is more likely to find a low priority cell to push out.

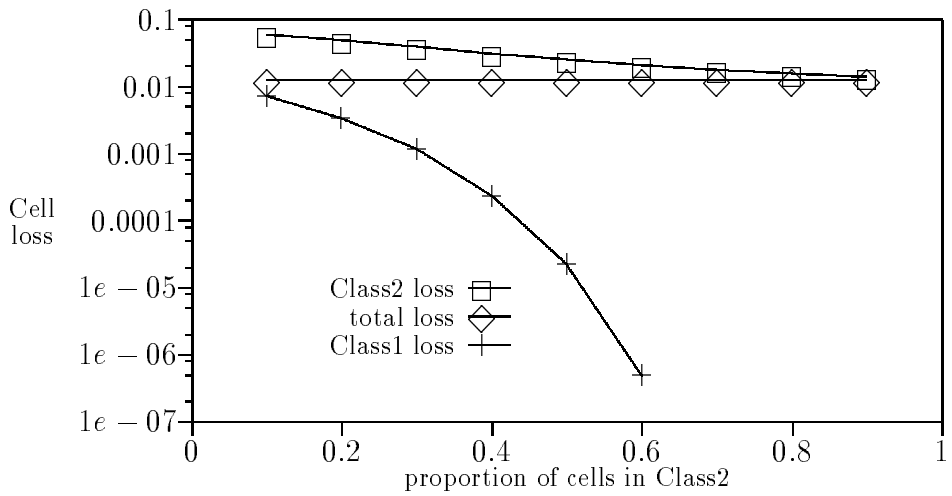

Figure 8: Voice cell loss for varying ratios of Class1/Class2 (load = 0.8, buffer size = 50)

5 Conclusions

Until extensive measurements can be made of traffic on a high bandwidth ATM network, studies of network performance must rely on estimates of the behaviour of applications on the network. This paper examines several traffic models that have been used to approximate network traffic in other environments with particular attention to their impact on

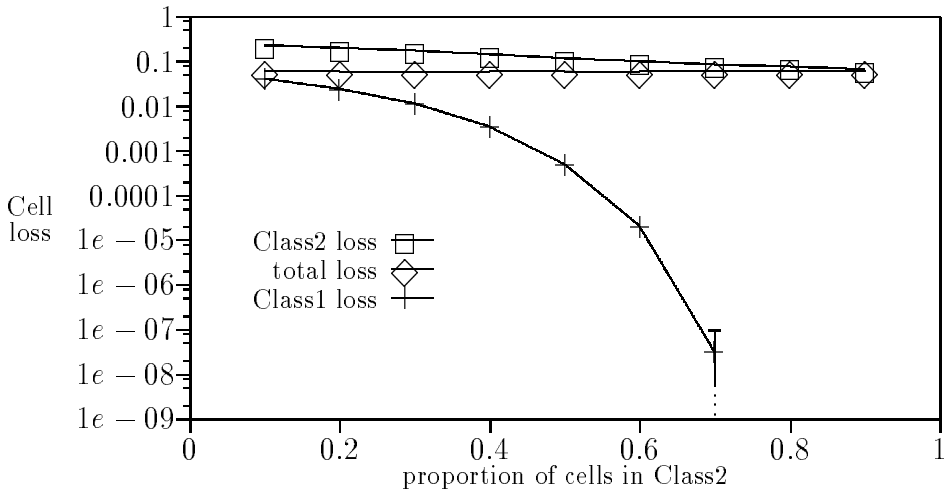

Figure 9: Data cell loss for varying ratios of Class1/Class2 (load = 0.8, buffer size = 50)

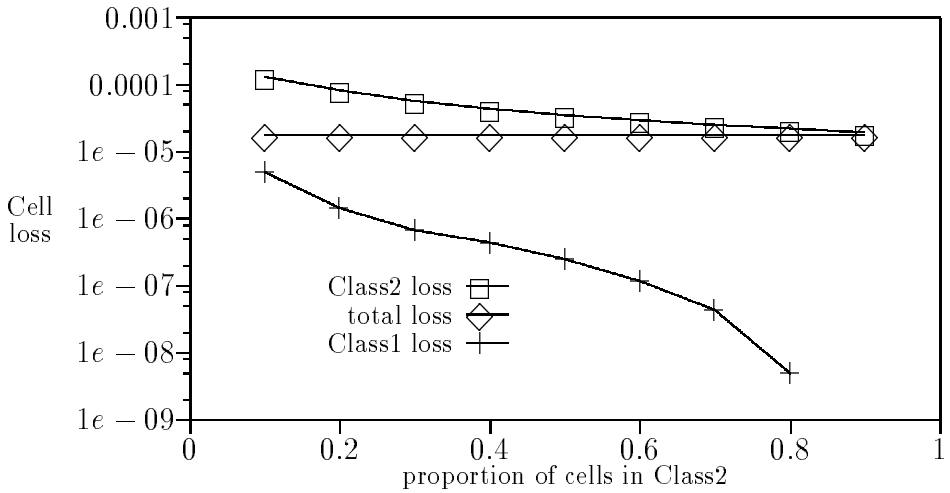

Figure 10: Video cell loss for varying ratios of Class1/Class2 (load = 0.8, buffer size = 50)

performance.

Previous work on priority queueing with Poisson traffic concluded that space priority queueing algorithms could effectively protect high priority traffic from excessive cell loss, even under strict QOS requirements. The range of cell loss anticipated by Poisson traffic was overly optimistic as the results of this paper establish. Real, bursty traffic will experience much higher cell loss than the Poisson traffic.

The results of this study demonstrate that space priority queueing can still provide good protection to high priority cells when the overload periods are not too long, the combination of the buffer size and load level lead to a reasonable total cell loss, and at least half of the cells have low priority. Priority queueing cannot stand alone as a congestion control mechanism, however. It depends on admission control algorithms and traffic shaping to keep the load at a manageable level. If the load on a switch becomes high for an extended period of time, the space priority queueing algorithms will be ineffective since the buffer will continually be full.

6 Acknowledgments

Mart Molle from the University of Toronto and Carey Williamson from the University of Saskatchewan read early drafts of this paper and provided many helpful comments. The research was supported by several organizations, including the Natural Sciences and Engineering Research Council of Canada, through Research Grant OGP0003707 and the Post-

graduate Scholarships program, and Telecommunications Research Laboratories (TR*Labs*).

References

[1] CCITT, Blue Book, Geneva, Switzerland. *CCITT Recommendation I.121, Broadband Aspects of ISDN*, 1989.

[2] G. Gallassi, G. Cappellini, L. Fratta, G. Rigolio, and F. Rossi. Performance evaluation of flow control mechanisms in multi-throughput class networks. In *Proceedings GLOBECOM 88*, pages 16.3.1–6, 1988.

[3] Jorge García and Olga Casals. Space priority mechanisms with bursty traffic. In *Performance of Distributed Systems and Integrated Communication Networks*, pages 393–412. Elsevier Science Publishers B.V., 1992.

[4] Ibrahim W. Habib and Tarek N. Saadawi. Multimedia traffic characteristics in broadband networks. *IEEE Communications Magazine*, pages 48–54, July 1992.

[5] Gérard Hebuterne and Annie Gravey. A space priority queueing mechanism for multiplexing ATM channels. *Computer Networks and ISDN Systems*, 20(1-5):37–44, December 1990.

[6] Harry Heffes and David M. Lucantoni. A markov modulated characterization of packetized voice and data traffic and related statistical multiplexer performance. *IEEE Journal on Selected Areas in Communications*, SAC-4(6):856–868, September 1986.

[7] Bjarne E. Helvik, Per Hokstad, and Norvald Stol. Correlation in ATM traffic streams – some results. In *ITC-13*, pages 25–32, 1991.

[8] Hans Kröner. Comparative study of space priority mechanisms for ATM networks. In *Proceedings IEEE INFOCOM '90*, pages 1136–1143, 1990.

[9] Arthur Y. M. Lin and John A. Silvester. Priority queueing strategies and buffer allocation protocols for traffic control at an ATM integrated broadband switching system. *IEEE Journal on Selected Areas in Communications*, 9(9):1524–1536, December 1991.

[10] Basil Maglaris, Dimitris Anastassiou, Prodip Sen, Gunnar Karlsson, and John D. Robbins. Performance models of statistical multiplexing in packet video communications. *IEEE Transactions on Communications*, 36(7):834–843, July 1988.

[11] K. L. Reid. *Priority Queueing in Asynchronous Transfer Mode Switches.* Master's thesis, Department of Computational Science, University of Saskatchewan, 1992.

[12] Hiroshi Saito. Queueing analysis of cell loss probability control in ATM networks. In *ITC-13*, pages 19–24, 1991.

[13] Gillian Woodruff and Rungroj Kositpaiboon. Multimedia traffic management principles for guaranteed ATM network performance. *IEEE Journal on Selected Areas in Communications*, 8(3):437–446, April 1990.