Traffic Shaping for Congestion Control in High Speed ATM Networks

Heng Liu
Richard B. Bunt
Carey L. Williamson
Department of Computational Science
University of Saskatchewan
Saskatoon, Canada

November 26, 1992

Abstract

A Broadband Integrated Services Digital Network (B-ISDN) will support a wide variety of services comprising voice, data and video information. The traffic generated by such services has radically different performance requirements and is of a very bursty nature. Such diverse requirements present new resource management challenges.

The bursty nature of the traffic is the primary cause of congestion in high speed networks – a phenomenon that results in performance degradation. Shaping the traffic as it enters the network makes it smoother (less bursty) and more predictable, thus reducing the probability of congestion. Traffic shaping is basically a conservative approach to congestion control that can protect the network from misbehaving sources when all traffic sources are properly shaped, but it may result in unused bandwidth.

This paper investigates the effects of a leaky bucket traffic shaper on the performance of an ATM-based B-ISDN network through simulation. The effect of various shaper parameters on performance is discussed. The results show that shaping can provide good network-level performance, but user-level (end-to-end) performance may be adversely affected.

1 Introduction

Packet-switched networks have changed considerably in recent years. The advent of high speed packet-switching and high reliability fiber-optic transmission has made it possible to develop a Broadband Integrated Services Digital Network (B-ISDN) – an all-purpose digital network that will support a wide variety of applications in a flexible, integrated, and cost-effective manner. The high transmission rate of B-ISDN, and the widely-varying service requirements of the traffic it carries, present new challenges for resource management, in particular congestion control.

The services supported in a multi-service network will have a wide range of performance requirements and a wide range of traffic flow characteristics. High-bit-rate traffic, which includes fixed-bit-rate or variable-bit-rate video and real-time services, requires a low delay and a guaranteed low cell-loss probability. Low-bit-rate traffic, including various types of data, does not have such stringent delay requirements but is more loss-sensitive.

The Asynchronous Transfer Mode (ATM) is a high-bandwidth, low-delay, packet switching and multiplexing technique that provides a unified means by which all traffic, connectionless as well as connection-oriented, can be transported on a common network [14]. A basic characteristic of packet-based transport is the sharing of network resources (processing resources, transmission bandwidth, and buffers) among a large number of information flows that are highly bursty and unpredictable. Since the traffic sources are bursty, the demands of simultaneous users may occasionally exceed the available network resources. Short overloads can be accommodated by the provision of buffers, which can absorb excess packets at the expense of delay. If the condition persists, however, congestion will occur and can lead to undesirable consequences such as heavy packet loss and long packet delay. Congestion control schemes try to protect the shared network resources from saturation by dynamically adjusting the traffic of the network.

Congestion control schemes can be classified into two categories: reactive control and preventive control [8]. With reactive control, sources adjust their traffic flows based on feedback information received from the network about the presence of congestion. With preventive control, sources must reserve network resources in advance, before they can access the network, and are required to remain within their allocated resources. Most congestion control mechanisms developed for conventional packet-switched networks are reactive approaches [10, 11, 16]. Senders adjust their traffic flows based on information about packet loss, delay, queue length, or other measureable performance criteria. Unfortunately, reactive approaches may encounter difficulties when applied to the high speed B-ISDN environment since propagation delay may dominate other delays in the network. By the time a source receives a feedback signal from the network to regulate its traffic, it may have already transmitted many packets (exacerbating the problem); or by the time it does react, the congestion may have already disappeared (rendering any subsequent action unnecessary). Preventive approaches are thus suggested for high speed networks to guarantee a certain performance level and reasonable network utilization.

Traffic shaping has been proposed to regulate the input rate and smooth the burstiness of traffic [2, 3, 4, 9]. A traffic shaper buffers the input traffic at the end systems and then releases it into the network at a controlled rate. Traffic shaping also acts as a bandwidth enforcement function. Shaping the traffic when it enters the network can make the traffic smoother, more predictable, and controllable, thereby reducing the probability of congestion inside the network. The performance gains due to traffic shaping are lower delay inside the network and lower probability of cell loss, but these gains come at the cost of extra delay introduced by traffic shaping before the traffic enters the network.

In this paper, the results of a detailed study of the effects of traffic shaping in an ATM network environment are presented. The paper focuses on the evaluation of tradeoffs associated with traffic shaping: the tradeoff between user-level performance and network-level performance; the tradeoff between trying to improve cell loss performance and the extra queueing delay at the shaper; and the tradeoff between buffer requirements at switches and buffer requirements at the shaper.

The remainder of the paper is organized as follows. Section 2 provides further motivation for traffic shaping. Section 3 describes the network model, the traffic model, and the traffic shaper model (a modifed leaky bucket scheme) that were used in the simulation study. Section 4 describes the experiments that were conducted with these models, experiments investigating the effects of traffic shaping on the performance of the network. The effects of various shaper parameters were also studied. The final section presents general conclusions concerning the merits and limitations of traffic shaping in a high speed network environment.

2 Traffic Shaping in an ATM-based B-ISDN

The purpose of B-ISDN is to provide an integrated services network supporting a wide range of digital traffic, from low-bit-rate traffic to high-bit-rate traffic. ATM offers the flexibility needed by B-ISDN. In an ATM network, bandwidth is divided into small and fixed-size packets, called *cells*, which are dynamically allocated to connections on demand. A cell consists of a 5-octet (byte) header followed by 48 octets of information. Routing of cells is based on the information contained in the header. Since cells belonging to a connection are identified by their header fields and there is no dedicated physical channel, the connection is only a virtual connection. In this way, bandwidth can be allocated to services on demand.

ATM networks are designed on a connection-oriented transport principle, which implies a setup of a virtual connection with end points. The request for the establishment of a connection is termed a *call*. Congestion control can be exercised at the call level and at the cell level [9]. At the call level, an admission control policy decides whether or not to accept an incoming call into the network. The purpose of admission control is to avoid long-term network congestion. During the call setup phase, network resources are checked and reserved along the connection route before the call starts. The call is accepted if network resources are available to satisfy the required class of service. Otherwise, it is rejected. At the cell level, bandwidth enforcement functions located in network access nodes monitor the traffic status. The bandwidth enforcement function regulates the input traffic rate to ensure that the traffic stays close to its allocated capacity so that it will not affect other connections in progress.

Most traffic sources are inherently bursty and unpredictable. When the discrepancy between assumed and actual traffic flow is significant, there will be a need to shape the traffic to conform to its allocated bandwidth. Traffic shaping helps to avoid congestion by controlling the rate at which cells are allowed to enter the network from a given source. A traffic shaper shapes the input traffic to match its claimed traffic descriptors: mean bit rate, peak bit rate and burst length. A traffic shaper acts like a filter, allowing cells that arrive smoothly to proceed into the network while delaying or dropping cells that violate the bandwidth "contract". Many schemes have been proposed and studied [8, 9], including the sliding window, the jumping window, the exponentially weighted moving average (EWMA), and the leaky bucket [18]. All these schemes shape the traffic for users, enforcing the contract for the network.

The *leaky bucket* mechanism is the most commonly proposed mechanism for traffic shaping [6, 17, 18]. In the leaky bucket approach, cells are dropped from traffic sources violating their bit rate contract. The basic leaky bucket algorithm enforces a mean bit rate and a maximum burst length by means of a token pool for which tokens are generated at a fixed rate. An arriving cell takes a token from the token pool and then goes through to the network. If no token is available, the cell is called a violation cell and is dropped immediately. The token generation rate thus enforces the mean rate at which cells are allowed to enter the network, while the size of the token pool limits the maximum length of cell bursts. This leaky bucket scheme can be modified [17] to provide input buffers for violation cells so as to better control the tradeoff between cell loss and queueing delay.

3 Description of Experiments

This section describes the experiments conducted to evaluate the effect of traffic shaping on the performance of the network, the results of which are presented in Section 4. A network model, a traffic model, and a traffic shaper moder were developed for these experiments. Traffic shaping is performed in each individual source prior to entering the network.

3.1 The Network Configuration Model

A simple network configuration model was developed for an ATM-based B-ISDN network. The model has four switches and connects ten source hosts (src) to ten destination hosts (dest), as shown in Figure 1. There is a traffic shaper for every traffic source. Each traffic source can represent a single call or an aggregation of a number of calls. Calls are routed deterministically to their destination hosts. Each link is a 150 Mb/s communication link, and all links are assumed to be error free. Ten virtual connections (src i to dest i, for $i = 1, 2, 3, \ldots, 10$) were provided, with path lengths of four switch hops (*i.e.*, crossing three switch-to-switch links). In the simulations considered for this paper all queueing and cell loss within the network occurs at the first switch. The effect of cross traffic at the other switches is studied in [13].

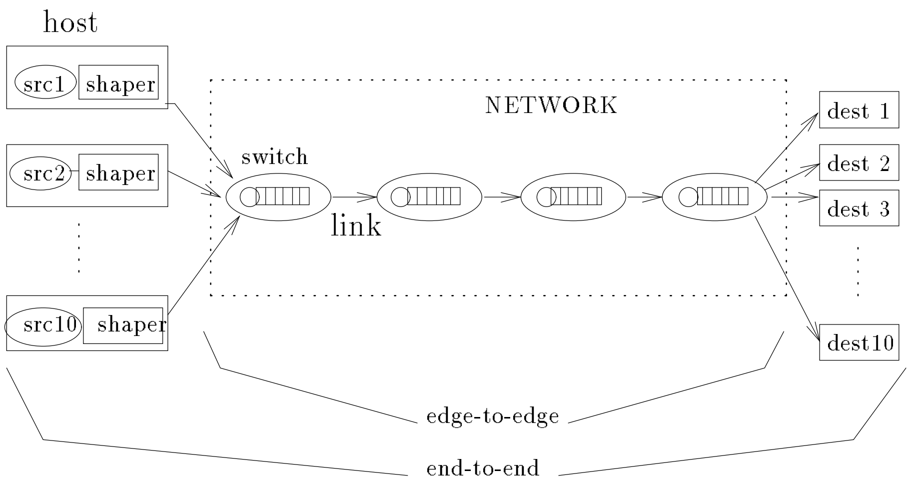

Figure 1: Network configuration for simulation

The model assumes internally non-blocking, output-buffered, fast packet switches. Output-buffered switches have been shown to perform much better than input-buffered switches and achieve optimal throughput-delay performance [1, 5]. For simplicity, a switching delay of zero is assumed. The service discipline in switches is FCFS.

The transmission time for a 53-byte ATM cell is approximately 2.8 μs on a 150 Mbps link. This transmission time is referred to as a *slot* and is the time unit used in the simulation. Link bandwidth is allocated in units of slots.

3.2 The Traffic Model

A two-state MMPP model was adopted for the traffic sources. The choice of MMPP was due to its popularity and its flexibility to model different traffic characteristics. The model is a Markov chain of two states as shown in Figure 2. Cells are generated according to a Poisson process in each state with mean arrival rates λ_1 and λ_2, respectively ($\lambda_1 > \lambda_2$). Holding times in each state are exponentially distributed with means $1/\beta$ and $1/\alpha$, respectively.

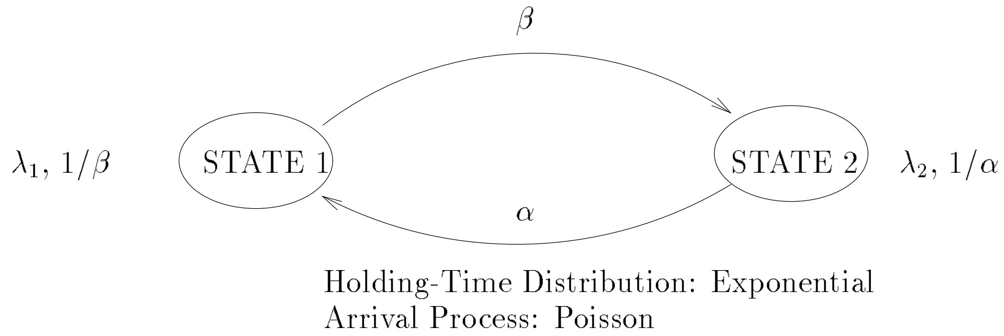

Figure 2: Two-state MMPP traffic model

In this paper, traffic is characterised in terms of its traffic descriptors. The following formulas define these quantities:

- mean bit rate: $B_m = (\alpha\lambda_1 + \beta\lambda_2)/(\alpha + \beta)$,
- peak bit rate: $B_p = \lambda_1$, and
- mean burst duration: $T = 1/\beta$.

Different levels of workload and different levels of burstiness can be simulated by varying the parameters of the model.

All traffic sources used in the simulations were modeled using this MMPP model. In all experiments reported in this paper, all sources have the same MMPP parameters, *i.e.*, the network is fed by homogeneous sources. Scenarios with heterogeneous traffic sources are considered in [13].

3.3 The Traffic Shaper Model

A traffic shaper employing a modified (buffered) leaky bucket scheme was used in this study. Its structure is shown in Figure 3.

The pool of tokens contains at most M tokens, the generation of which is deterministic. At a token generation rate of r tokens/slot, one new token arrives every $1/r$ slots. The token generation rate determines the mean rate at which cells enter the network. An arriving cell enters a finite buffer first. If a cell in the shaper buffer finds the token pool non-empty, it takes a token and enters the network immediately. If the shaper buffer is full, the cell is dropped. The size of the shaper buffer (denoted as A) represents a tradeoff between cell loss and cell delay in the shaper. The larger the

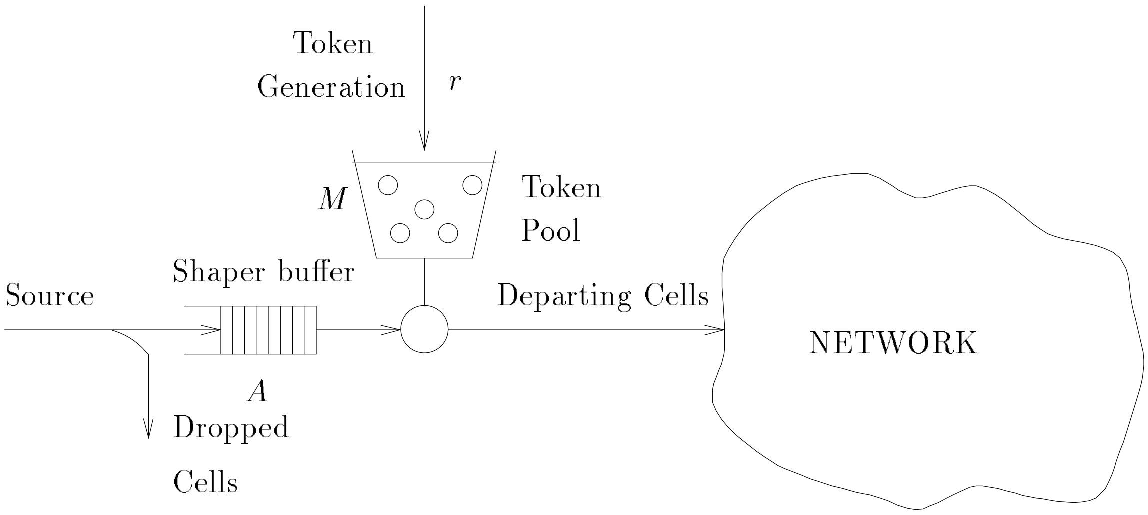

Figure 3: Traffic shaper model

shaper buffer, the lower is the probability of cell loss but the higher is the maximum queueing delay.

3.4 Performance Measures

Two basic criteria are used in the evaluation of overall performance of the network: cell loss and queueing delay. Cell loss is measured as the fraction of cells generated at the sources that fail to arrive at their respective destinations. Cells may be dropped at the shaper or at any switch within the network. Queueing delay is measured as the time cells spend in bufers en route to their destinations. Again, cells may be queued at the shaper or at any switch within the network.

Performance can be assessed at the network-level (edge-to-edge performance) or at the user-level (end-to-end performance), as shown in Figure 1. Network-level performance is measured from the network entry access point to the exit access point. This reflects cell loss and queueing delay that connections experience inside the network. User-level performance is measured from the time of a cell's generation at the source host (src) until its delivery to the destination host (dest). The user-level performance measures reflect cell loss and queueing delay incurred by connections before their cells enter the network (*i.e.*, at the shaper) plus that incurred after they enter the network (*i.e.*, including network-level performance). These two levels of performance may differ since the presence of the traffic shaper may introduce extra cell loss and delay. Both levels of performance were studied, as was done in [12].

4 Simulation Results and Analysis

4.1 Parameter Settings

Traffic on the network can be characterised by the following parameters:

- the burstiness level of the traffic b, defined as the ratio of the peak rate to the mean rate ($b = B_p/B_m$),
- the load level ρ, defined as the sum of the mean rates of all sources, and
- the mean burst duration T, defined as the average time spent in the active state.

These experiments focussed on the performance of the network at heavy load (*i.e.*, ρ in the range from 0.8 to 1) since the purpose of the investigation was to study networks under congestion. So that traffic was sufficiently bursty, burstiness levels ranged from a low of 2 to a high of 10. When the load level is high, a burstiness level of 10 can give a very high peak cell rate (such as 0.95 cells/slot for a single source). The mean burst duration T varied from a low of 10 time slots to a high of 100 time slots. Traffic with a very long burst duration (e.g. 1000 cells) cannot obtain reasonable performance with switch buffers of the order of 100 cells [15], a figure commonly used in ATM research [7]. Buffer size ranged from a low of 10 cells to a high of 600 cells in these experiments.

The general effects of traffic shaping are considered first, followed by the specific impact of different shaper parameters: the shaper buffer size A, the token pool size M, and the token generation rate r. These parameters determine the characteristics of the traffic after shaping, and hence determine the performance, both network-level and user-level. The impact of traffic shaping on the requirement for buffers at switches is studied. The conservative nature of traffic shaping is examined as well. To obtain statistically meaningful results, sample sizes for random variables and simulation run lengths were carefully chosen throughout the experiments. The confidence interval for all experiments was $\pm 1\%$ of the mean value of the performance measure at a 95% level.

4.2 General Effects of Traffic Shaping

Ten homogeneous sources were used to investigate the general effects of traffic shaping on network performance. Each source had a burst duration of $T = 10$ and a burstiness level of $b = 10$. The aggregate load level increased from 0.8 to 1.0. The switch buffer size was fixed at $B = 60$. All traffic sources were subjected to traffic shaping. The traffic shaper used a buffer size $A = 60$, and a token pool size $M = 20$. Each shaper generated tokens at a rate $r = 0.11$ tokens/slot representing an admission rate of one cell every 9 slots from each source.

Figure 4 shows the effect of traffic shaping on cell loss. For comparison purposes, the performance without traffic shaping is also included. With traffic shaping, cell loss inside the network is reduced substantially. The difference between the "user-level" curve and the "network-level" curve is the result of cell loss at the shaper. Buffering at the shaper can reduce this loss. Traffic shaping, with or without a buffer, provides better loss performance than no shaping at all.

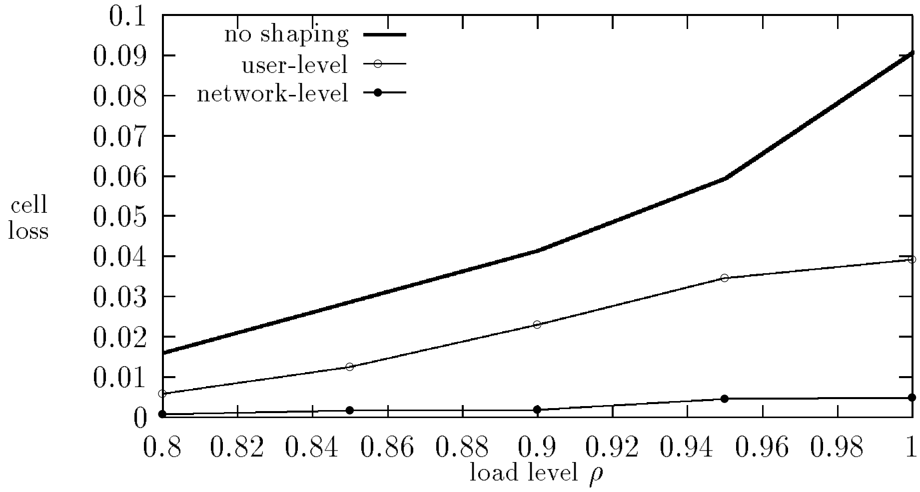

Figure 4: The effect of traffic shaping on cell loss

Buffering cells to reduce loss comes at the expense of additional delay. Figure 5 shows the effect of traffic shaping on queueing delay. Since the traffic is very bursty, many cells arriving in a burst are forced to wait at the shaper before they are eligible to enter the network. Delays will be accumulated for cells generated towards the end of each burst since the token pool is usually empty then. The shaper buffer gradually clears between bursts. From the user-level perspective, traffic experiences a higher delay than without traffic shaping. Much of this delay happens in the shaper. From the network-level perspective, since traffic shaping provides a network environment where all traffic has been smoothed, cells experience little delay inside the network.

These results demonstrate that traffic shaping is very desirable for loss-sensitive traffic, such as data transfer, since it can reduce cell loss substantially. The extra delay, however, makes it less desirable for delay-sensitive traffic, such as real-time video, since this traffic may be unable to tolerate the extra delay introduced by traffic shaping.

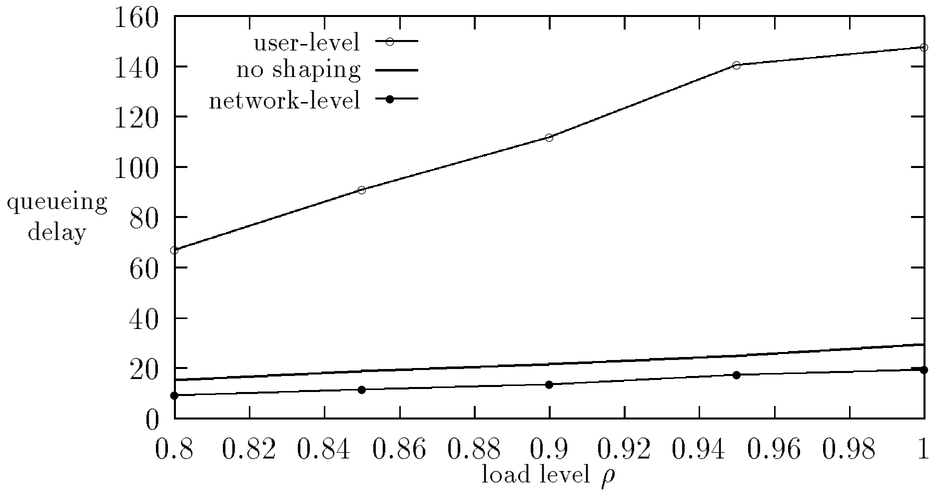

Figure 5: The effect of traffic shaping on queueing delay

4.3 Effects of Traffic Shaper Parameters

Shaper parameters determine the characteristics of traffic after shaping. This in turn determines the performance of the network both at the user level and at the network level.

The shaper buffer size A controls the maximum number of cells waiting for access to the network at any given time. It provides a tradeoff between cell loss at the shaper and waiting time at the shaper. When a long burst of cells arrives, tokens saved in the token pool will soon be used up, and all subsequent cells are either dropped or, if there is room, placed in the shaper buffer to enter the network at the token generation rate. For a given cell placed in the shaper buffer, its waiting time is a function of the number of cells waiting before it in the buffer and the interval between token generations. If the token generation rate is 0.1, tokens are generated at an interval of 10 time slots. A cell waiting at the end of a buffer which has a size of $A = 100$ will have to wait for 1000 time slots to enter the network! This observation suggests that a small shaper buffer might be more suitable.

The tradeoff between loss and delay at the shaper was studied in more detail. The token pool size at the shaper was fixed at $M = 10$ and the token generation rate was fixed at $r = 0.11$. The size of the shaper buffer A was varied from 20 to 120 in steps of 20. Figures 6 and 7 give the results of this experiment. Figure 6 shows the effect of shaper buffer size A on cell loss. As A increases, user-level cell loss drops quickly at first, then begins to flatten out. Increasing the shaper buffer size has a big impact on loss performance at first, as cells that might otherwise have been dropped at the shaper are buffered instead. Beyond a certain point the effect becomes quite limited.

For example, increasing A from 20 to 40 reduces cell loss from 14.7% to 7.6%, while increasing A from 100 to 120 reduces cell loss from 1.7% to 1.1%.

From the network-level perspective, since traffic is smoothed, it experiences little cell loss inside the network. As A increases, fewer cells are dropped at the shaper, meaning more cells are allowed to enter the network. So there is a slight increase in network-level cell loss as A increases.

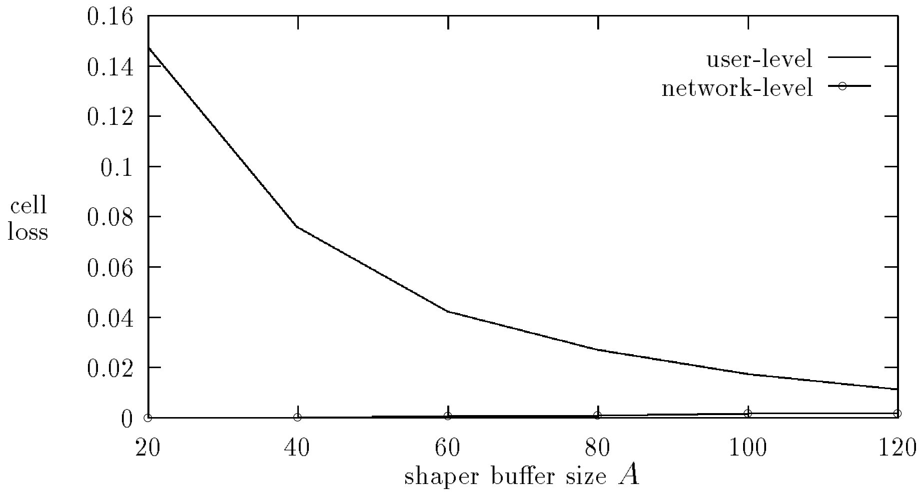

Figure 6: The effect of the shaper buffer size on cell loss

Figure 7 shows how the shaper buffer size A affects queueing delay. User-level queueing delay increases rapidly as A increases, while network-level queueing delay remains fairly constant. That means that most of the increase in queueing delay occurs at the shaper. Since cells are slowed down at the shaper and enter the network smoothly, they experience little delay once they are inside the network. The size of the buffer at the shaper, therefore, has little effect on network-level delay.

These results demonstrate that a shaper buffer reduces cell loss at the cost of high queueing delay at the shaper. The choice of A determines the tradeoff between this cell loss and the queueing delay. If A is too big, it will result in an unacceptable queueing delay at the shaper. If A is too small, however, cell loss at the shaper will be too high. Choosing a suitable A depends on the characteristics of the traffic source such as its burstiness level and burst duration, and the other shaper parameters as well. A detailed study of the effect of these parameters is presented in [13].

4.4 Token Pool Size

The size of the token pool M determines how many tokens can be "saved up" for a given traffic source, and thus the maximum number of cells that can be injected into

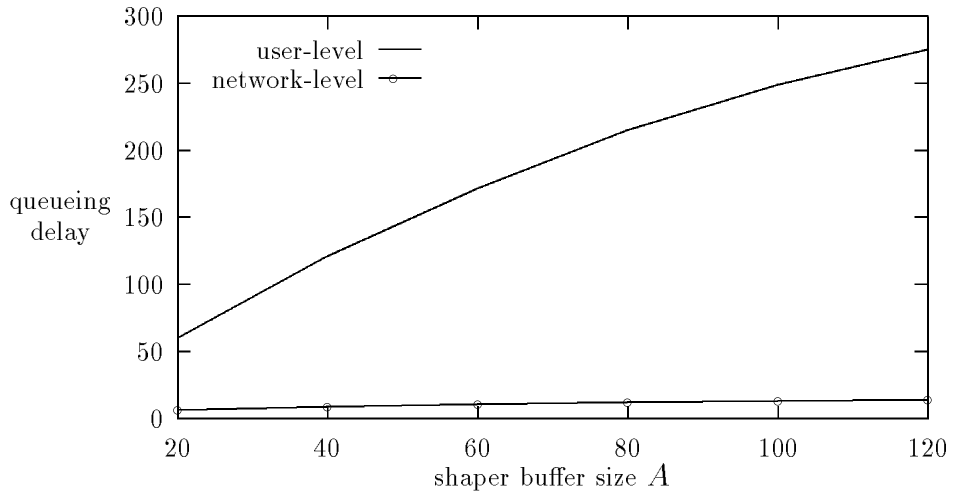

Figure 7: The effect of the shaper buffer size on queueing delay

the network back-to-back at a time. The larger the token pool size, the more cells can enter the network at a burst, so the burstier is the traffic. This subsection presents the results of an experiment conducted to study the effect of M on performance.

For this experiment, the shaper buffer size was fixed at 60 and the token generation rate was fixed at 0.11. The token pool size was varied from 10 to 80 in steps of 10.

Figure 8 shows the effect of M on cell loss. Increasing M results in fewer cells lost in the shaper. Allowing more cells to enter the network in a burst, however, obviously leads to more cells lost inside the network. The larger M is, the bigger is the burst that is permitted to enter the network, and so the higher is the network-level cell loss.

Figure 9 shows how the token pool size M affects the queueing delay. When M increases, more tokens can be saved in the token pool, so cells arriving in bursts may spend less time waiting at the shaper, hence user-level queueing delay decreases. M has little impact on the network-level queueing delay since this is determined primarily by the switch buffer size.

4.5 Token Generation Rate

The token generation rate of the traffic shaper, r, determines the mean rate at which cells enter the network. If the token generation rate is lower than the mean input cell rate, many cells will be dropped in the shaper. The token generation rate also represents the proportion of bandwidth allocated to the call, *i.e.*, the token generation rate is set corresponding to the bandwidth allocated to the source. For bursty traffic, if the call reserves bandwidth just equal to its mean rate it may lose many cells when they come in a burst. A bursty traffic source requires a token generation rate

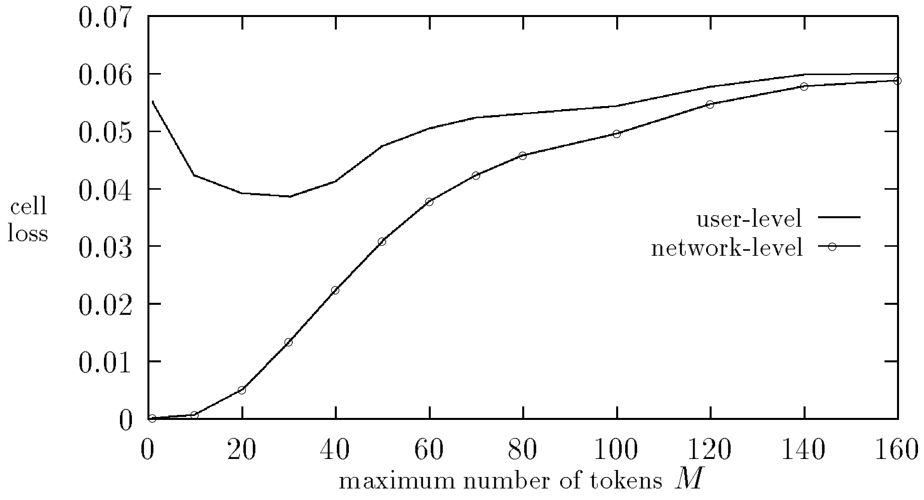

Figure 8: The effect of token pool size on cell loss

higher than its mean cell rate, *i.e.*, an amount of bandwidth higher than its mean rate. Since the traffic does not actually use all the bandwidth allocated to it, the network may allow some over-reservation of the bandwidth in the sense that the sum of the token generation rates of all sources can be higher than 1. But too much over-reservation may cause the performance of the network to deteriorate. In this subsection, the results of an experiment that studied the effect of the parameter r on overall performance are presented.

For the traffic shaper, the shaper buffer size was fixed at $A = 80$ and the token pool size was fixed at $M = 10$. The token generation rate r was varied from slightly below the mean cell arrival rate to twice the mean rate, that is, from $r = 0.095$ to $r = 0.19$ (the sum of all r's thus ranged from 0.95 to 1.9). Bandwidth is over-reserved from the point $r = 0.1$, *i.e.*, when the ratio of the token rate over the mean rate is greater than 1.0.

Figures 10 and 11 show the loss and delay performance, respectively, as a function of r. For comparison, the performance of the case where there is no traffic shaping is included and indicated as "no-shaping".

When the token generation rate is equal to the mean rate of the input traffic, user-level cell loss and queueing delay are high (*e.g.*, cell loss of 8.9% and queueing delay of 375 slots). For bursty traffic sources, many cells arrive in bursts and at a rate much higher than the source's mean rate (*i.e.*, the peak cell rate). If the token generation rate is equal to the mean cell rate, cells in a burst will soon use up all the tokens in the token pool, and the remaining cells will have to wait at the shaper buffer and enter the network at the token generation rate. The excess cells that the shaper buffer cannot handle will be dropped. This result demonstrates that making

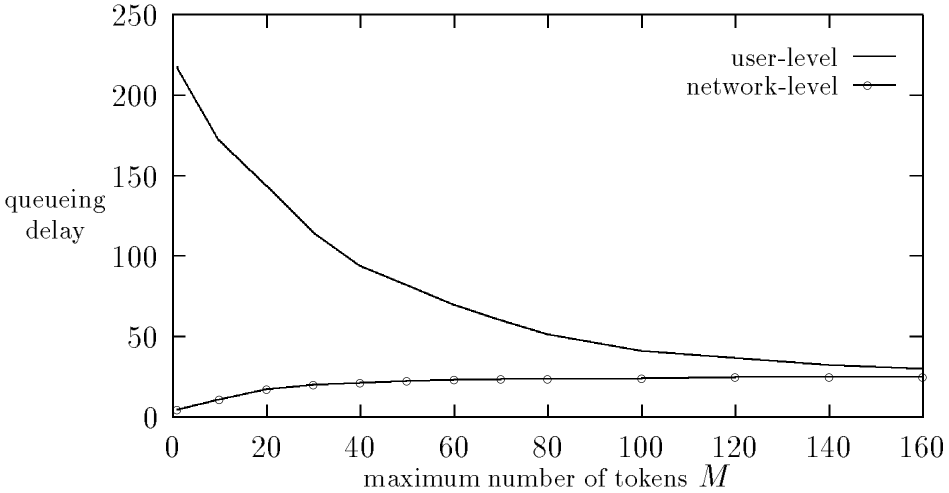

Figure 9: The effect of token pool size on queueing delay

r equal to the mean rate is not a good choice.

Increasing r means more bandwidth is allocated to the call. The larger r is, the faster tokens are generated, and so fewer cells are lost at the shaper and cells spend less time waiting at the shaper. Although fewer cells are lost at the shaper, more cells begin to be dropped inside the network. The total cell loss begins to increase when r passes a certain point since too many cells are lost inside the network. When r is very big, no cells are dropped at the shaper. All cells dropped are dropped inside the network. This does not mean, however, that traffic shaping has no effect on performance. Since the token generation rate is still less than the peak rate, the shaper is still effective in smoothing the bursts. The total cell loss is still lower than in the case where there is no traffic shaping. As far as queueing delay goes, user-level delay drops quickly as r increases. Network-level delay increases slightly as r increases, since more cells enter the network. As r continues to increase, the user-level and network-level delay curves begin to converge.

These results demonstrate that bursty traffic must reserve bandwidth a little higher than its mean rate. It also shows that network bandwidth can be over-reserved to a certain extent. For the traffic source discussed here, a token rate that is about 1.2 times the mean rate seems to be a suitable choice. In this case, the sum of the token generation rate of all sources is 1.14, which implies that the bandwidth is over-reserved by 14%.

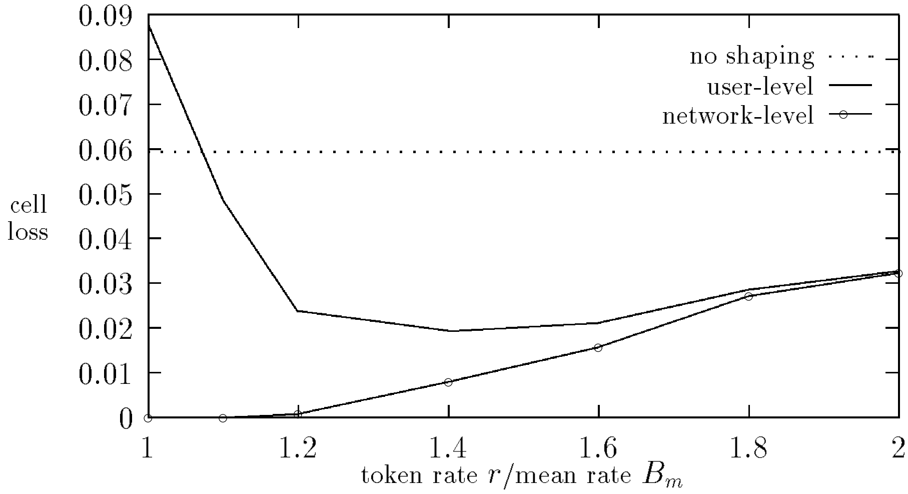

Figure 10: The effect of token generation rate on cell loss

4.6 Buffer Requirements at Switches

Traffic shaping buffers excessive cells at the shaper instead of at the switches, *i.e.*, outside the network instead of inside the network. Since traffic shaping spaces out bursts of cells before they enter the network, traffic inside the network is very smooth, hence less buffer space is required at the switches. An experiment was performed to study the effect of traffic shaping on the buffer requirements at the switches. Each source had a burst duration of 10 time slots, the burstiness level was fixed at $b = 10$, and the load level was kept constant at $\rho = 0.95$. The traffic shapers of all sources had the same set of parameters: $A = 60$, $M = 10$, and $r = 0.11$.

The results show that traffic shaping reduces the requirement for buffer space at the switches. For example, 100 cells of buffer space at the switch results in a cell loss probability of approximately 3% without traffic shaping (see Figure 12). If the traffic is shaped, cells are buffered at the shaper, so the same level of cell loss can be obtained with only 15 buffers at the switch.

4.7 The Conservative Nature of Traffic Shaping

In essence, traffic shaping slows down the source traffic to force it enter the network more gracefully. In some cases, however, the network may be able to absorb long bursts of cells from one or more sources without becoming severely congested. For example, a long burst of cells from one source might happen to come to the network when all other sources are in their idle states. Without traffic shaping, such a burst could get through the network since there would be enough excess bandwidth to accommodate it. With traffic shaping, however, the burst will be delayed at the

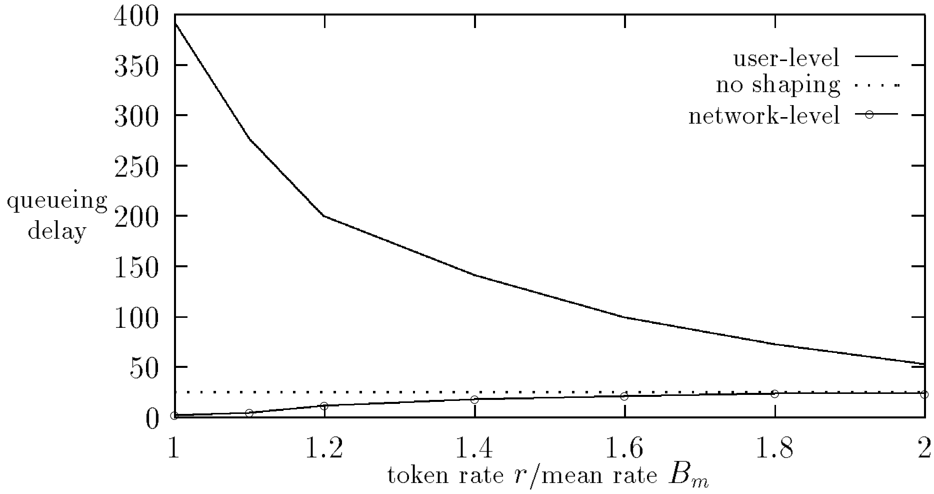

Figure 11: The effect of token generation rate on delay

shaper and sent at the token generation rate or even be lost. In this sense, the traffic shaper may be too conservative in its approach.

Figure 13 shows cell loss performance as a function of burst duration, both with shaping and without. At first, when the burst duration is short, cell loss performance with traffic shaping, both from the user-level perspective and from the network-level perspective, is better than without shaping. As the burst duration increases, user-level cell loss increases faster with shaping than without shaping. The two curves cross at around $T = 40$, beyond which point traffic shaping results in poorer performance than no shaping since almost all losses occur at the shaper.

This result demonstrates that while traffic shaping limits the rate at which cells are allowed to enter the network, it also limits the opportunity to exploit statistical gain when multiplexing bursty traffic sources. In this respect traffic shaping may be too conservative an approach to congestion control.

5 Conclusions

This paper describes the results of a series of simulation experiments conducted to study the performance impact of a leaky bucket traffic shaper in a high speed ATM network. The results presented demonstrate the effect of the various shaper parameters on performance and lead to the following general conclusions.

Traffic shaping is an effective technique in some respects. In particular, it provides good network-level performance, *i.e.*, low cell loss and low queueing delay inside the network. It may have an adverse impact on user-level performance, however, since the

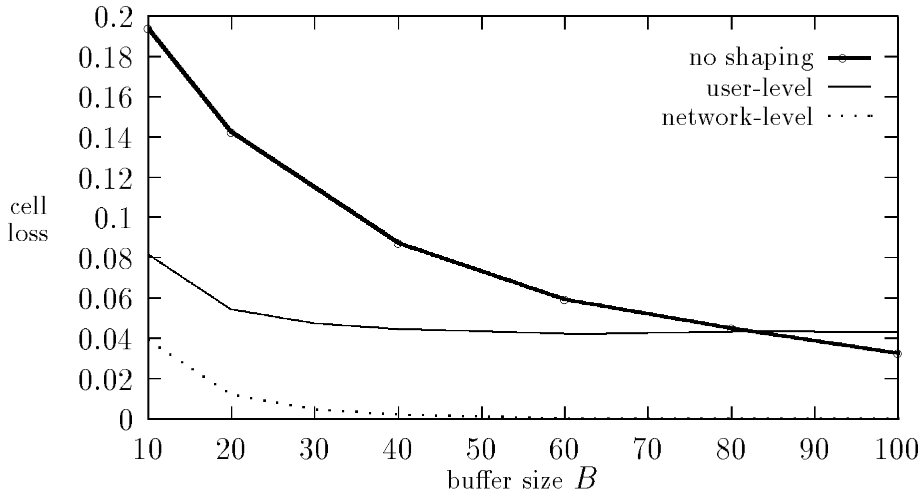

Figure 12: Cell loss performance with different switch buffer sizes

good network-level performance comes at the expense of extra cell loss and queueing delay at the shaper. Overall, although traffic shaping can reduce cell loss substantially, it does so at the cost of extra delay. This suggests that traffic shaping might be more attractive for loss-sensitive traffic than for delay-sensitive traffic.

Traffic shaping provides protection to the network from the demands of misbehaving users since performance degradation can be limited to those users. In this way it is more feasible to provide performance guarantees to network users when sources are subjected to traffic shaping.

Finally, as a technique for congestion control, traffic shaping is a very conservative approach. Although it can limit the undesirable effects of congestion it may result in wasted bandwidth in some situations. This cost must, of course, be weighed against the advantages of traffic shaping in the context of any particular network.

6 Acknowledgements

This research was supported, in part, by the Natural Sciences and Engineering Research Council of Canada (NSERC) under Research Grants OGP0003707 and OGP0121969 and Strategic Grant STR0045250. The work benefitted considerably from regular discussions with members of the DISCUS Research Group at the University of Saskatchewan and from the technical support provided by Darryl Willick.

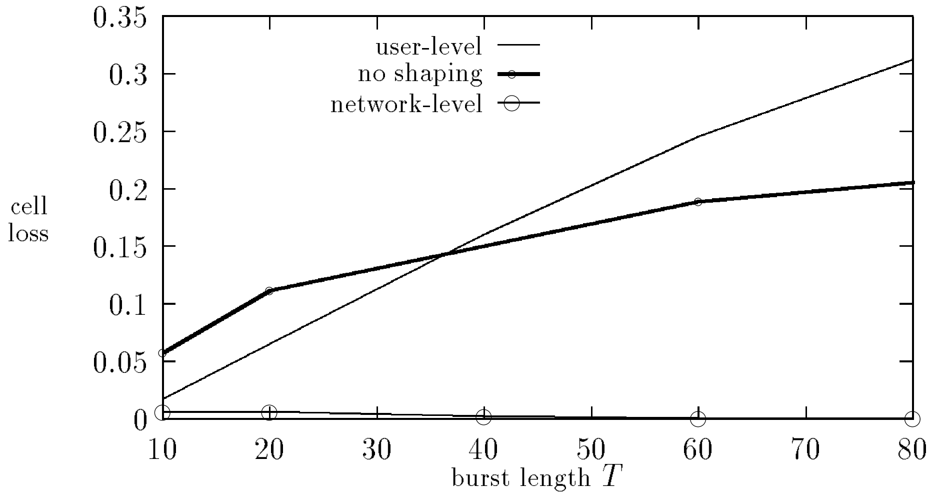

Figure 13: The effect of burst duration on cell loss

References

[1] H. Ahmadi and W. Denzel, "A Survey of Modern High-Performance Switching Techniques", *IEEE Journal on Selected Areas in Communication*, Vol. SAC-7, No. 7, Sept. 1989, pp. 1091-1103.

[2] A. Berger, "Performance Analysis of a Rate Control Throttle Where Tokens and Jobs Queue", *Proceedings of INFOCOMM 1990*, June 1990, pp. 30-38.

[3] M. Butto, E. Cavallero and A. Tonietti, "Effectiveness of the 'Leaky Bucket' Policing Mechanism in ATM Networks", *IEEE Journal on Selected Areas in Communications*, Vol. 9, No.3, April 1991, pp. 335-342.

[4] I. Cidon, I. Gopal and R. Guerin, "Bandwidth Management and Congestion Control in plaNET", *IEEE Communications Magazine*, Oct. 1991, pp. 54-64.

[5] G. Daddis and H. Torng, "A Taxonomy of Broadband Integrated Switching Architectures", *IEEE Communications Magazine*, May 1989, pp. 32-42.

[6] A. Eckberg, D. Luan and D. Lucantoni, "Bandwidth Management: A Congestion Control Strategy for Broadband Packet Networks – Characterizing the Throughput-Burstiness Filter", *Computer Networks and ISDN Systems*, Vol. 20, 1990, pp. 415-423.

[7] V. Friesen, *A Performance Study of Broadband Networks*, M.Math. Thesis, Dept. of Computer Science, University of Waterloo, 1991.

[8] H. Gilbert, O. Aboul-Magd and V. Phung, "Developing a Cohesive Traffic Management Strategy for ATM Networks", *IEEE Communications Magazine*, Oct. 1991, pp. 36-45.

[9] I. Habib and T. Saadawi, "Controlling Flow and Avoiding Congestion in Broadband Networks", *IEEE Communications Magazine*, Oct. 1991, pp. 46-53.

[10] V. Jacobson, "Congestion Avoidance and Control", *Proceedings of ACM SIGCOMM '88*, Stanford, California, Aug. 1988, pp. 314-329.

[11] R. Jain, "Congestion Control in Computer Networks: Issues and Trends", *IEEE Network*, Vol. 4, No. 3, May 1990, pp. 24-30.

[12] W. Leland, "Window-Based Congestion Management in Broadband ATM Networks: The Performance of Three Access Control Policies", *Proceedings of INFOCOMM 1989*, Ottawa, Canada, 1989, pp. 49.7.1-49.7.7.

[13] H. Liu, *Traffic Shaping for Congestion Control in High Speed ATM Networks*, M.Sc. Thesis, Dept. of Computational Science, University of Saskatchewan, 1992.

[14] S. Minzer, "Broadband ISDN and Asynchronous Transfer Mode (ATM)", *IEEE Communications Magazine*, May 1989, pp. 17-24.

[15] M. Murata, Y. Oie, T. Suda and H. Miyahara, "Analysis of a Discrete-Time Single-Server Queue with Burst Inputs for Traffic Control in ATM Networks", *IEEE Journal on Selected Areas in Communication*, Vol. SAC-8, No. 3, April 1990, pp. 447-458.

[16] K. Ramakrishnan and R. Jain, "A Binary Feedback Scheme for Congestion Avoidance in Computer Networks with a Connectionless Network Layer", *Proceedings of ACM SIGCOMM '88*, Stanford, California, Aug. 1988, pp. 303-313.

[17] M. Sidi, W. Liu, I. Cidon and I. Gopal, "Congestion Control Through Input Rate Regulation", *Proceedings of INFOCOMM 1989*, Ottawa, Canada, pp. 49.2.1-49.2.5.

[18] J. Turner, "New Direction in Communications (or Which Way to the Information Age?)", *IEEE Communications Magazine*, Vol. 24, No. 10, 1986, pp. 8-15.