Cell Delay Variation and Burst Expansion in ATM Networks: Results from a Practical Study Using Fairisle

Simon Crosby and Ian Leslie,
University of Cambridge Computer Laboratory *

Peter Key
Transport Design and Performance Unit, BT Laboratories †

October 25, 1995

Abstract

Current interest in cell delay variation in ATM networks arises from the need to provide guarantees on loss and delay for traffic with real time constraints. This study uses the Fairisle ATM LAN to measure the CDV experienced by CBR and periodic, bursty sources when multiplexed with Bernoulli, VBR video and LAN interconnection traffic, in a network of non-blocking ATM switches.

We find that higher bitrate connections experience greater relative CDV for the same total load, and that larger ATM switches cause greater CDV. The Bernoulli model is a poor predictor of the performance using real traffic. The CDV experienced by a CBR connection increases over successive switching stages, but is bounded above by an asymptote which depends on the source bitrate, the total load, and the type of background traffic.

Bursts become separated by background traffic, causing a decrease in CDV. We study burst expansion by tracing individual cells in each burst, to determine the expansion rate, and relate the results to two analytical approximations.

Finally, we study the performance of reference streams under static priority and rate-based scheduling, concluding that significant performance benefits can be achieved using non-FCFS queueing policies.

1 Introduction

Cell delay variation (CDV) is a measure of how closely the temporal characteristics of a stream of ATM cells match those of an ideally spaced, constant bit rate (CBR) cell stream of the same mean rate. It is intimately related to the leaky bucket algorithm, proposed by the ITU-T and ATM Forum as a technique for policing the peak rate of traffic entering the network, and is also useful for determining how much a CBR stream has been perturbed by interfering traffic. The algorithm relies on observation of the cell stream at a single point[1], predicting the arrival time of the next cell from a connection based on that of the current

*email: {Simon.Crosby, Ian.Leslie}@cl.cam.ac.uk phone: +44 1223 334600 fax: +44 1223 334678
†email: pbk@perf.bt.co.uk phone: +44 1473 645820 fax: +44 1473 645846
[1]This is also known as the 1-point CDV measure

cell and the negotiated peak rate. Cells which arrive "early" may be discarded, subject to a tolerance parameter.

The causes of CDV are well known: queueing delays and the effects of switching and multiplexing can cause cells from a connection to become grouped together, artificially increasing the peak transmission rate and making the traffic more bursty. This, in turn, makes it difficult for switches in the network to provide guarantees on cell delay without discarding traffic (alternatively, more buffer space is needed to ensure that loss constraints are met). It has even been suggested that connections should be policed at inter-network boundaries to drop cells which exceed the negotiated peak rate. This will affect the performance experienced by the user, and will complicate the traffic controls within the network. It is important that network control mechanisms ensure that users who comply with their traffic contracts are not penalised as a result of the (hidden) effects of multiplexing. Our study aims to determine whether such controls are necessary and what parameters critically affect the CDV sustained by traffic of different types.

We also study a process known as burst expansion – the mechanism whereby bursts of cells entering an ATM network tend to become smoothed by mixing with other traffic. An understanding of burst expansion is important for several reasons: badly behaved users who transmit bursts of traffic into the network could impact the QoS experienced by other connections. In addition, the CDV of a bursty source gives us an indication of the effects of a policing mechanism, for example at an inter-network boundary, on a well behaved CBR connection which has become more bursty as a result of multiplexing.

To date, most research into CDV in ATM networks has been based on analytical models and simulation studies [17, 12, 15, 9]. In contrast we report on the findings of a practical study using a real ATM network, and relate our findings to the analytical predictions. Recently the RACE EXPLOIT project has reported the practical validation of several analytical models of CDV [4, 14]. This work suffers from several drawbacks: first, the traffic is artificially generated to conform to the assumptions of the analytical models. Second, to achieve high loads the output traffic from a switch must be re-introduced at another input, causing correlation in the traffic mix. Finally, the use of commercial ATM equipment limits the ability to measure important features of the cell scale process. The ATM network used in this study is the Fairisle ATM LAN, which provides hardware support for high time-resolution measurements of network activity. Rather than rely on hypothetical traffic processes which may not be accurate, we have used sources of real ATM video and data traffic. We consider the effects of the following factors on the CDV experienced by a reference connection:

- the *sizes of the switches* (number of ports) in the network which the connection traverses,
- the *types of background traffic* with which the reference source contends for buffer space and transmission capacity,
- the *number of switching stages* traversed by the connection,
- the total *network load* on each link traversed by the connection,
- the *proportion of the total load* contributed by the the connection under study, and
- the *queueing discipline* implemented by the switches on the path of the connection.

We study two types of reference sources, namely CBR sources, which have a constant inter-cell time, and periodic bursty sources which are parameterised by a constant burst size and a constant inter-burst time.

2 The Fairisle ATM Network

The Fairisle project has developed an ATM platform to permit research into performance, architectural and management issues in ATM networks [3]. It is based on a network architecture composed of ATM switches which are interconnected to form a general topology ATM LAN. The Fairisle switch is based on a simple, 4×4 input buffered crossbar. Switches of degree greater than 4 are constructed by using multiple 4×4 crossbars in a 2-stage delta interconnection network.

The hardware design of the Fairisle network components permits a high degree of experimental flexibility owing to the use of field programmable gate arrays which implement a number of features to support experimental work, including high resolution clocks and hardware time stamping of cells. Fairisle can be configured as either an input or output buffered switch. Each port is equipped with 2K cell buffers, 4 MB of RAM and an ARM RISC processor, which implements all policy-specific traffic control functions, using a flexible control plane which exports RPC-style interfaces [2, 19] for signalling and management. This has enabled us to implement a wide range of measurement facilities to accurately monitor network performance, and to experiment with non-FCFS queueing disciplines which attempt to provide guarantees to different types of traffic.

3 Experimental Network Configuration

Many manufacturers of ATM equipment base their switch designs on an effectively non-blocking and hence output buffered switching fabric. Although this increases the complexity and cost of the switch, output buffered designs are appealing because they alleviate the effects of contention due to switching. In a non-blocking switch one cell from each input can be delivered to a single given output in a single switching cycle.

In our study each experiment aims to determine the CDV experienced by a single reference connection in a network of output buffered ATM switches. Although we have performed experiments using Fairisle in an input buffered configuration, the results will be reported in a separate publication.

To construct an output buffered switch using Fairisle we need to ensure that the switching stage is non-blocking. This is accomplished by using two input buffered Fairisle switches in tandem to implement a single output buffered switch, in the configuration shown in Figure 1. The abbreviation BGT is used for "background traffic", and REF for "reference source".

In the diagram each link is labelled with its transmission rate, relative to the full line rate, which is denoted as rate 1. The transmission rates of the sources are scaled by $1/D$, where D is the degree of the switch. The transmission link rate for the output ports of the first switch is set to the full link rate, 1. Traffic arriving at the inputs of the switch is switched by the first fabric and multiplexed onto the full rate link. The output link of the second switch is set to a rate $1/D$. Traffic arriving at the input of the second switch experiences blocking due

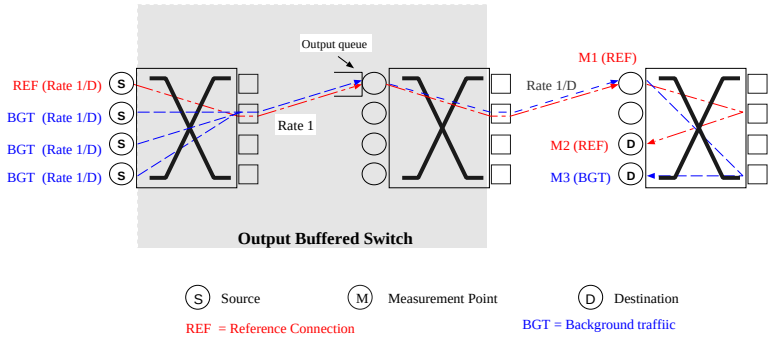


Figure 1: Single stage non-blocking (output buffered) switching configuration

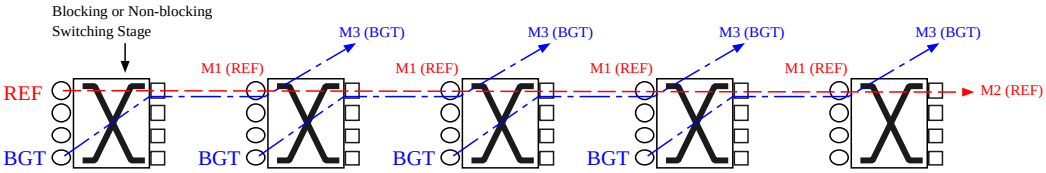


Figure 2: Network topology for multiple switching stages

to back pressure across the fabric from the output running at the reduced rate. Queueing occurs at this point, which can be viewed as the output buffer of the switch so constructed. The output buffered switch has a switching capacity of D times the input and output port transmission rates, and is therefore non-blocking.

3.1 Measurement Points

Figure 1 shows some of the measurement points implemented to obtain performance results for each experiment:

- M1 – measures the CDV experienced by cells from the reference source, and their inter-arrival time distribution.
- M2 – measures the offered load from the reference source, and traces the arrival times of the first 10^6 cells.
- M3 – measures the offered load of the background traffic to ensure experimental accuracy.

The accuracy of the experimental load was monitored by measuring the contributed load from the reference stream and the background traffic at independent measuring points. The load from each source and the total multiplexed load is accurate to within less than 1%. All measurements were made using a clock which has a resolution of about 1/10 of the cell slot time.

The network topology for experiments involving more than one stage of switching is shown in Figure 2. At each switch the background traffic departs and a new (independent) source of background traffic of the same type is multiplexed with the reference stream.

4 Traffic Sources

In each experiment a single reference connection contends with one of several types of background traffic for buffer space and bandwidth. We limit the buffer space available for queueing to be 2000 cells, ensuring that the experiments are loss free. We use trace-driven traffic generators as sources, using long, independent traces measured from actual sources. We can also record a trace of the arrival times of cells from the reference connection after 4 switching stages. This can be used as the input reference source to a second experiment, allowing us to arbitrarily extend the logical path length. In this way we have extended our experiments to 12 stages of switching.

4.1 Reference Traffic Sources

Our principal concern in these experiments was to investigate the effects of different types of background traffic on reference streams of two types: CBR streams and periodic, bursty streams. The former are representative of synchronous services adapted onto an ATM bearer, for example an E1 (2 Mb/s) trunk. The latter permit us to study the effects of background traffic on bursts as they proceed through the network. In both cases we used the same basic traffic parameters – the bursty reference connections have the same mean rate as their CBR counterparts of the same name, but transmit in bursts of a fixed size. Section 6 presents results for experiments using CBR sources, and Section 7 using bursty sources. Table 1 shows for each source its name, its (AAL) bitrate, the number of cells it transmits per second (on a 100 Mb/s TAXI link), its inter-cell emission period T (in slots, rounded to the nearest integer) and the load ρ_{ref} it places on the link.

Reference Source	Bitrate (Mb/s)	Cells/s	T	ρ_{ref}
B	2	5461.3	41	0.024
E	10	27306.7	8	0.120

Table 1: Reference source parameters

4.2 Background Traffic

Key parameters of the background traffic which affect the CDV experienced by a reference connection include its peak and mean rates. The ratio of the two is sometimes called the "burstiness" of the traffic. Below we specify these parameters for each of the 3 types of background traffic used.

Bernoulli Traffic

A Bernoulli traffic source is the discrete time analogue of a Poisson process, and is frequently used in analytical models of ATM multiplexers. We used Bernoulli traffic to facilitate comparison of our results with the predictions of analytical models and simulations, although it is not representative of any particular ATM traffic. In each experiment the *total* load ρ_{tot} on each link was one of 3 values: 0.5, 0.8 or 0.95. The load placed on the link by the Bernoulli traffic is given by $\rho_b = \rho_{\text{tot}} - \rho_{\text{ref}}$, where ρ_{ref} is the load placed on the link by the reference connection.

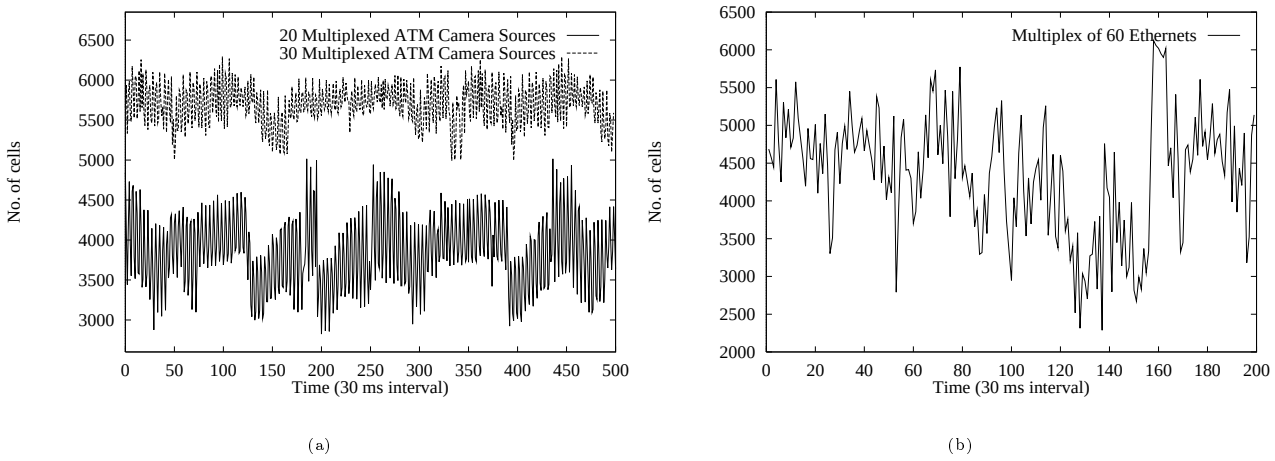


Figure 3: Measured load for multiplexed (a) ATM Camera and (b) Ethernet sources

ATM Camera Traffic

Our video source is the Fore Systems AVA 200 ATM camera, which can transmit video of various formats and qualities directly into the ATM network. To construct the sources used in these experiments we superimposed several pre-recorded traces of ATM camera video to form a multiplexed load representing 20 and 30 cameras. The camera was set up to transmit 25 frames per second, JPEG compressed, 24 bits per pixel colour video from normal BBC TV transmission, with a picture size of 48×33 tiles (each 8×8 pixels); each tile is compressed and transmitted with a single ATM cell of overhead. The amount of information per tile differs according to the video content. We used an early version of the camera without a spacing policer, hence back-to-back cells were occasionally transmitted. The peak rate of each source is thus 100 Mb/s and the mean rate approximately 2.5 Mb/s, giving a burstiness of about 40. The traces include a wide range of scenes. The total mean load with each source is given in Table 2. Figure 3 (a) shows a section of a trace of the measured load of a multiplex of 20 and 30 ATM camera sources. Each point on the graph is the number of cells which arrived during that 30 ms interval.

Reference Source	ρ_{tot} (20 sources)	ρ_{tot} (30 sources)
B	0.58	0.83
E	0.68	0.93

Table 2: AVA-200 ATM Camera traffic parameters

Data Traffic

A substantial portion of WAN ATM traffic is likely to be "LAN interconnection" traffic linking remote LAN sites. Such traffic is packet-based and may be extremely bursty. Our data sources were constructed from the Bellcore Ethernet traces [13]. We created two sources from 50 and 60 independent segments of the traces; each Ethernet segment, representing a single LAN site, was peak rate limited to half the link rate to reduce its burstiness. Without this the aggregate

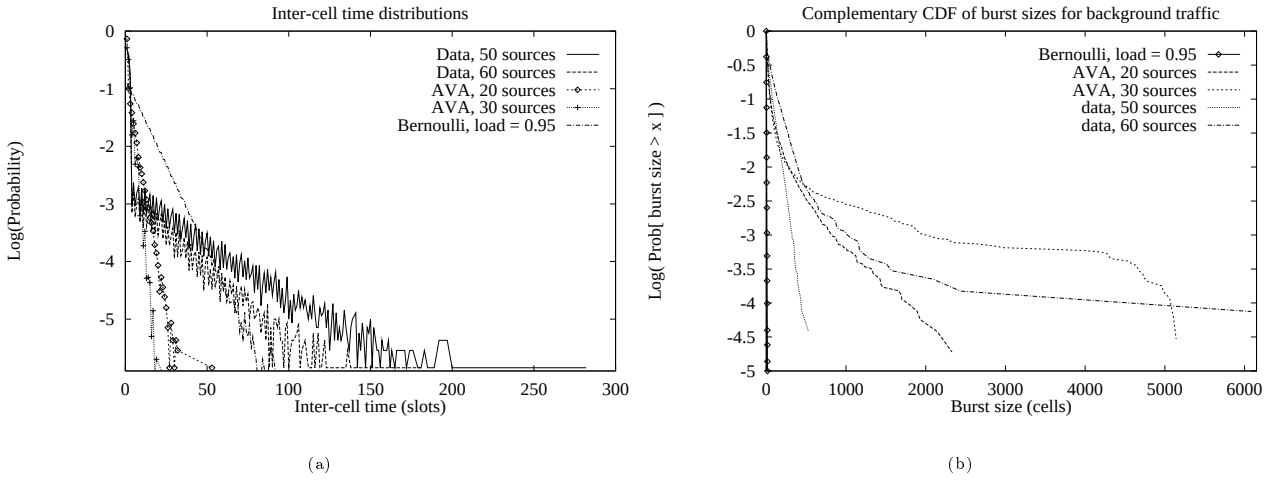


Figure 4: Inter-cell time distributions and complementary CDF of burst sizes for background traffic

traffic is too bursty for the available buffer space. The two traces we use differ significantly – the trace with 60 Ethernets has bigger bursts and a higher mean load per segment, causing the disproportionate increase in load shown in Table 3. The average burstiness of a single segment is 60.1 (50 segments) and 48.7 (60 segments). Figure 3 (b) shows the measured load for a short period of the data traffic composed from 60 multiplexed Ethernet segments. The graph shows extreme burstiness of this type of background traffic. Although much wide area traffic will be TCP-based and therefore sensitive to network congestion, experimentation with a large number of dynamically controlled sources was not possible.

Reference Source	ρ_{tot} (50 sources)	ρ_{tot} (60 sources)
B	0.44	0.64
E	0.55	0.75

Table 3: LAN interconnection traffic parameters

Summary

Figure 4 (a) shows the measured properties of the inter-cell time distributions for each type of background traffic. Of particular interest is the shape of the distributions for the data traffic: a large number of cells arrive back-to-back or close together (packets of traffic), and there is a long exponential tail. Cumulative Distribution Function (CDF) of the burst size distribution for each type is shown in Figure 4 (b); here a burst is defined to be a group of cells separated by 5 or more cell times. The graph shows clearly the effects of multiplexing bursty sources – the burst size distributions have extremely long tails compared to Bernoulli traffic, and point to potential performance problems in the network.

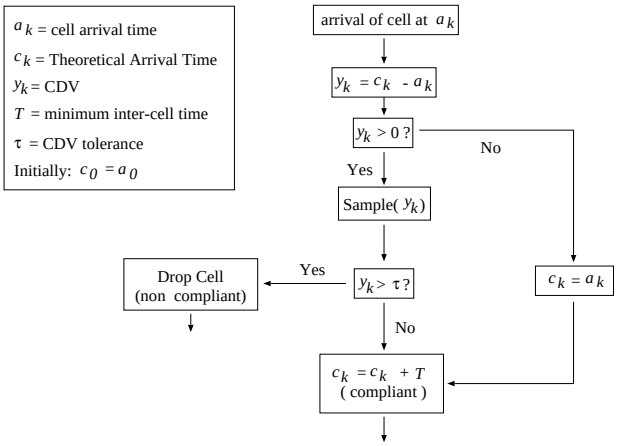


Figure 5: The Sampling I.371 Leaky Bucket Policer

5 Statistical Measures

In each experiment the reference connection was monitored at each multiplexing point and at the sink. The results for experiments using up to 4 switches are based on the transmission of 10^7 cells from the reference connection, whereas those from 4 to 12 stages are based on only 10^6 cells. The latter aimed only to study trends in the development of CDV beyond 4 stages, but the results seem to be accurate in spite of the relatively small sample size.

The I.371 Leaky Bucket

In these experiments the CDV was measured using a version of the ITU-T Recommendation I.371 leaky bucket algorithm 5 which samples each positive CDV value, shown in Figure. For a given reference source s the leaky bucket was parameterised with T_s, the minimum inter-cell time of the source, and the selected tolerance τ_s to accommodate CDV introduced by multiplexing. A value $\tau_s = \infty$ was used, so no cells were dropped. To do this it was necessary to speed up the leaky bucket to ensure that the CDV remained bounded – a value of $T = 0.95T_s$ was used, a 5% speedup in the leak rate.

The I.371 definition of CDV is also called the 1-point CDV measure as it relies on observation from a single point in the network. A problem with the 1-point CDV measure is its dependence on the time at which measurement commences: given the same input stream, but starting at different points, the algorithm will produce a different sequence of observed CDV values, and if used to police a reference connection will discard different cells in each case. It can be shown, however that as the length of the observation period tends to infinity, the *proportion* of cells discarded is the same irrespective of the starting point. An alternative, exponentially weighted measure of the CDV has been proposed, which produces a unique CDV distribution irrespective of the starting point [18]. A drawback of this algorithm is the complexity of performing the exponential weighting on-the-fly. An alternative measure, the 2-point CDV, is defined to be the difference between the absolute cell transfer delay between two measurement points and the delay of a reference cell which experiences no contention. It is difficult to measure in a network without a (synchronous) global clock, and its utility as a statistical measure is limited [9].

VARIANCE IN INTER-CELL TIMES

To gain a better statistical understanding of the traffic process we also measured the inter-cell time distribution of the reference source at the exit of each switch. This allows us to compare our results with an analytical approximation due to van den Berg and Resing [20] and extended by Kelly and Key [12], which we reproduce below. The analysis shows that a CBR connection traversing successive stages of switching where it is multiplexed with Poisson traffic at high load, asymptotically approaches a maximum perturbation. Let X_n be the the inter-cell departure time of a given pair of reference cells from the n^{th} multiplexer. The process $(X_n, n = 0 \ldots \infty)$ is a Markov chain from which expressions for the mean and variance of X_n can be derived. In particular, after n switching stages the variance of the inter-cell time distribution is given by

$$\mathrm{Var(X_n)} = \frac{\mathrm{T^2(T-1)}}{\mathrm{2T-1}} \left(1 - \left(\frac{\mathrm{T-1}}{\mathrm{T}} \right)^{\mathrm{2n}} \right). \tag{1}$$

It is clear from Equation (1) that the convergence of $\mathrm{Var(X_n)}$ is slow if T, the source period, is large. Thus a low bitrate source should take longer to reach its maximum perturbation. Also, the absolute value of $\mathrm{Var(X_n)}$ is greater for a reference connection with a larger value of T – thus a low bitrate source should experience a greater variance in inter-cell times than a high bitrate source. $\mathrm{Var(X_n)}$ differs significantly from the I.371 measure of CDV which is dependent on the the order in which cells arrive and only records cells which are "early" i.e. which exceed the negotiated peak rate. We will attempt to relate the two measures, on the basis of measured results.

6 Results and Discussion – CBR Sources

In this section we investigate the influence of several variables on the CDV experienced by CBR sources. For each source s the results are presented in units of T_s, the source period. To convert them into link slot times simply multiply by the value of T_s given in Table 1.

6.1 Total Load

Figure 6 (a) shows the effect of increasing Bernoulli load on the two reference sources, for a single stage of switching. The increase in CDV with load is non-linear (this is confirmed by the results which follow), although each curve itself is approximately log linear where sufficient data points occur. Although the CDV experienced by source E is worse than that experienced by source B (in units of T), the absolute values of the CDV (in slot times) experienced by source B and E are similar at high load.

6.2 Switch Size

Figure 6 (b) investigates the effect of switch size on CDV. An increase in switch size increases the CDV experienced by the reference source. An increase is to be expected given that a reference source will see more variability in the offered load across the switching fabric when the number of ports increases.

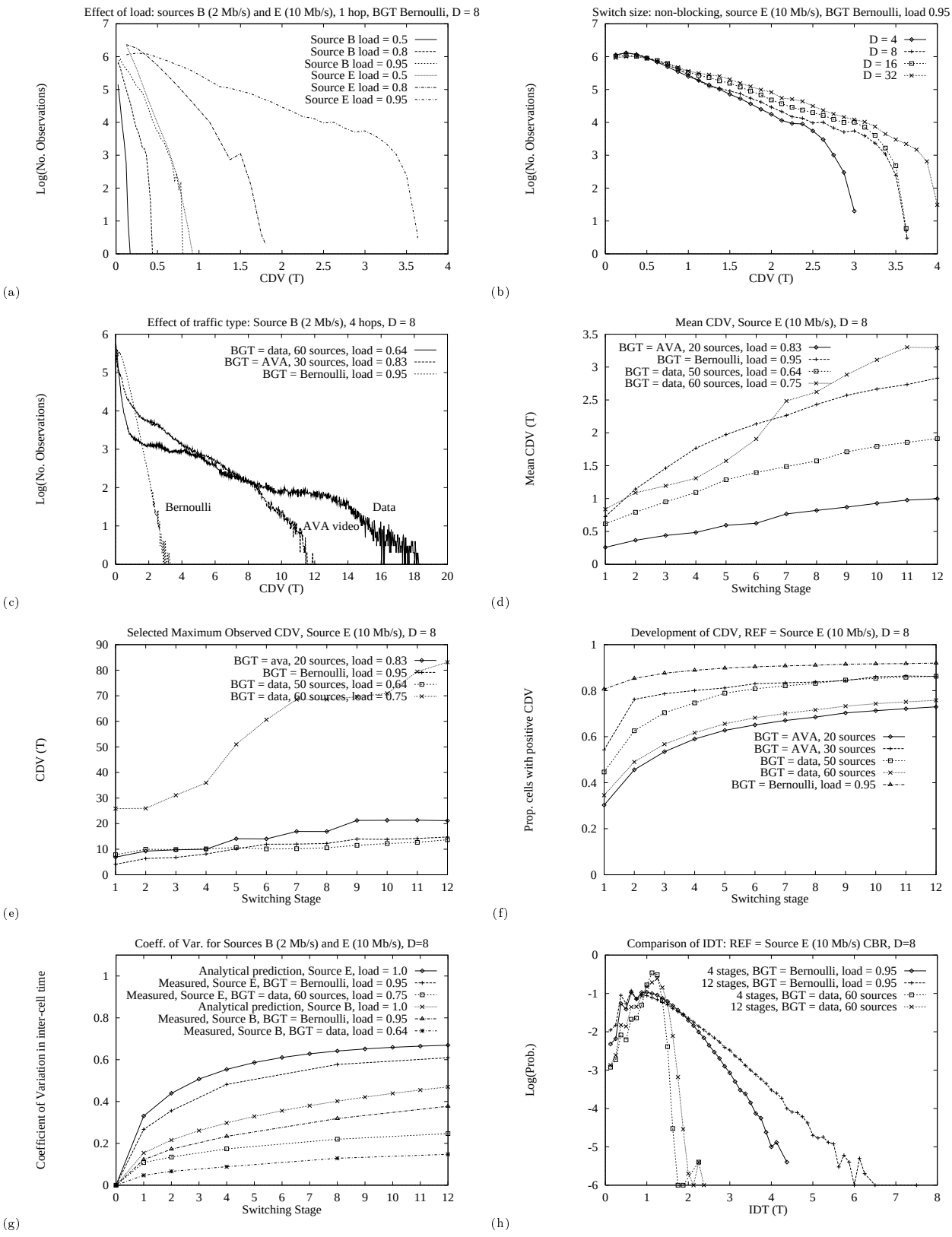


Figure 6: CDV results for CBR sources

6.3 Background Traffic Type

Figure 6 (c) shows the enormous difference in CDV experienced by source B using different background traffic types. The graph shows the CDV after 4 stages of switching, and illustrates the inability of the Bernoulli model to accurately predict the CDV using real traffic, in spite of its higher load. With data and video background traffic the CDV is badly affected by their burstiness, with a long tail and uneven distribution. This suggests caution when using Bernoulli models for performance predictions in ATM networks.

6.4 Number of Switching Stages

As a CBR connection proceeds through multiple stages of switching, its CDV increases. We would like to be able to predict, for a given source, the expected CDV with a particular type of background traffic, after a number of switching stages. Comparison of the tail value of the CDV for source B and a load of 0.95 between Figures 6 (a) and (c) shows that the CDV increases markedly from 1 to 4 stages. Figures 6 (d), (e) and (f) quantify the change in CDV for source E, for 12 stages of switching, by plotting in (d) the mean CDV[2] in (e) the extreme value of the CDV, and in (f) the proportion of cells in the experimental sample which experienced positive CDV. The mean develops smoothly: it appears to be bounded above by a value which is a function of the background traffic type and total load. The extreme CDV value behaves similarly, except for the high loads of the bursty traffic types. The proportion of cells which suffer positive CDV (and which are therefore candidates for discard, subject to the tolerance τ) also grows in a bounded fashion. After 12 stages of switching a high percentage (80-90%) of the total sample has experienced CDV, with Bernoulli traffic inflicting the most damage. Interestingly the low load of data traffic causes a higher proportion of cells to experience positive CDV than the high load of data traffic. Note also that the rate of increase in the number of cells with positive cdv is directly related to the burstiness of the background traffic, with Bernoulli traffic yielding an almost constant value and the low load of video traffic showing an extremely rapid increase.

We conclude that the maximum CDV appears to be bounded above irrespective of the background traffic type. For low total loads the increase in CDV with additional switching stages is small, whilst for high loads the increase is far greater. For lower bitrate reference sources the rate of increase in CDV is lower and takes more switching stages to stabilize.

The I.371 leaky bucket is difficult to model analytically. To understand the process causing the development of the CDV we turn to the approximation described in Section 5. Figure 6 (g) compares the measured coefficient of variation in the inter-cell departure time from each switching stage and the predicted coefficient of variation from the approximation for the inter-cell time variance described by Equation (1), for the two CBR sources. The model assumes Poisson background traffic and a load of 1. The measured results using Bernoulli background traffic with a load of 0.95 compare well with the predicted values, for both source B and E.

The measured results confirm the prediction that the variance is bounded above by a value which is a function of the total load and the type of the background traffic, as well as the source rate $1/T$. Since the variance in inter-cell times is greater for sources whose period T is larger, we conclude that lower bitrate sources require *larger* playout buffers at the sink to

[2]The CDV using 30 AVA sources as background traffic is far higher and is not shown here.

ensure the same loss probability.

For bursty background traffic, the variance (and coefficient of variation) is lower. This seems somewhat counter-intuitive, but can be explained as follows: using bursty background traffic a reference cell arriving at a queue will tend to find it either relatively idle, in which case it will not be delayed, or heavily loaded by a burst of background traffic, in which case additional reference cells have a higher chance of "catching up" with the delayed cell. If the bursts of background traffic are large and infrequent the latter case will be more common, and thus the reference traffic will tend to form bursts of cells close together.

This hypothesis is borne out by Figure 6 (h) which compares the log of the IDT distributions after 4 and 12 stages of switching, for source B using Bernoulli and data background traffic. The tail of the IDT distribution using data traffic is extremely short, and hence the variance of the distribution is lower.

7 Burst Expansion

The expansion of bursts of ATM traffic due to the effects of mixing is of great interest because the rate at which bursts expand determines the degree of damage they can inflict on other traffic. In addition, the CDV of a bursty reference source gives us an indication of the effects of an I.371 leaky bucket policer on a well behaved CBR connection which, as a result of multiplexing with other traffic, has become more bursty. In this section we investigate burst expansion using Fairisle, relating our results to two analytical models.

7.1 The Burst Expansion Rate

The COST 224 project has suggested an analytical approximation for burst expansion which aims to approximate the rate at which a burst of traffic will expand as it proceeds through the network [17]. Consider a periodic burst, the cells of which arrive at the multiplexer at time instants $S_i = iD, i = 0 \ldots M$, where D is the switch size. During a burst, a cell from the burst will arrive at the switch every D slots. Assume that other arrivals to the multiplexer constitute a Poisson process with parameter ρ, and let $\lambda = 1/D$. Let $V(t)$ be the amount of unfinished work at time t. Now $V(S_0-)$ is distributed as the virtual waiting time in an $M/D/1$ queue with load ρ whereas $V(S_M-)$ is distributed almost exactly by that of an $M + D/D/1$ queue, just before the deterministic arrival of a CBR cell. As the burst passes through additional multiplexers its inter-arrival times become more random and so the unfinished work seen by the last cell can be estimated as that of an $M/D/1$ queue with load $\rho + \lambda$. The burst expands as the cells become separated by the background traffic. The expected increase in the total time spanned by the burst, from time the first cell leaves a switch until the last cell leaves the switch, is called the expected increase in burst size $\mathrm{E}(\Delta\mathrm{B})$, expressed in slots per stage of switching. The pessimistic estimation for the expected increase of the burst size is given by

$$\mathrm{E}(\Delta\mathrm{B}) \approx \frac{\lambda}{2(1-\rho)(1-(\rho+\lambda))}. \tag{2}$$

For example, with $D = 20, \rho = 0.9$ and $\rho + 1/D = 0.95$ the expected burst expansion is $\mathrm{E}(\Delta\mathrm{B}) \approx 5$.

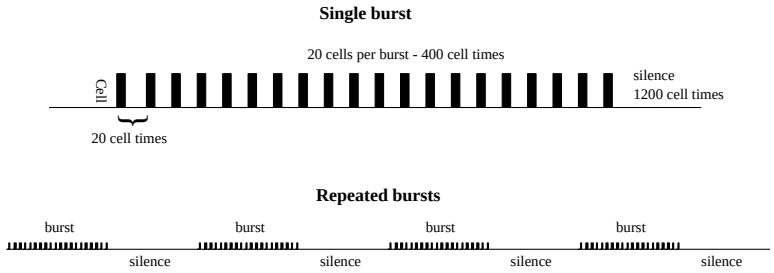


Figure 7: Graphical representation of the COST 224 reference source

In [17] this approximation was compared to simulated results for a network with 100 switching stages using the above parameters. The simulation results show an approximately linearly proceeding expansion and suggest that the bound in Equation (2) overestimates the expansion by a factor of two.

We have replicated the COST 224 experiment, using the same parameters: a switch of size $D = 20$, background Bernoulli traffic of load $\rho_b = 0.9$ and a periodic bursty reference source with a burst size of 20 cells. Within the burst cells are emitted with a period of 20 cell times (this reference source is also referred to as the 1 in 20 source). We set the inter-burst gap to be 1200 cell times, to allow the $Geo/D/1$ queue to stabilize between bursts. The total load during a burst is thus $\rho + 1/D = 0.95$. Figure 7 shows the bursty reference source graphically, on two time scales: a single burst, and multiple bursts separated by a long silence period.

Figure 8 (a) shows an example of the mean burst expansion over the 12 switching stages for each of the 20 cells within a burst (the first cell is the closest to the x axis). The cells enter the network spaced apart by T. As they proceed through the network they are separated by the Bernoulli background traffic. The data were obtained by recording at each multiplexer the inter-arrival time of every cell pair, and the mean was calculated from 10000 successive bursts.

Background traffic	$\rho + 1/D$	E(ΔB) (slots)
Poisson	0.95	5.00 (analytical)
Poisson	0.95	2.69 (simulation)
Bernoulli	0.95	4.36 (measured)
AVA video, 30 sources	0.86	2.09 (measured)
Bellcore data, 60 sources	0.68	1.99 (measured)

Table 4: Comparison of COST 224 and experimental results

From the experimental results we can measure the mean burst expansion, for Bernoulli and other background traffic types. We compare these to the analytical and simulation results in Table 4, which shows the close match between the measured, analytical and simulation results for Bernoulli traffic. The expansion is lower for the video and data background traffic, both because bursts of reference traffic have a higher probability of being unmodified by bursty background traffic, and also because these sources have a lower mean load.

We have extended the above results by studying the same source but with the cells in a burst grouped back-to-back. We also used bursty versions of sources B and E. In each case each

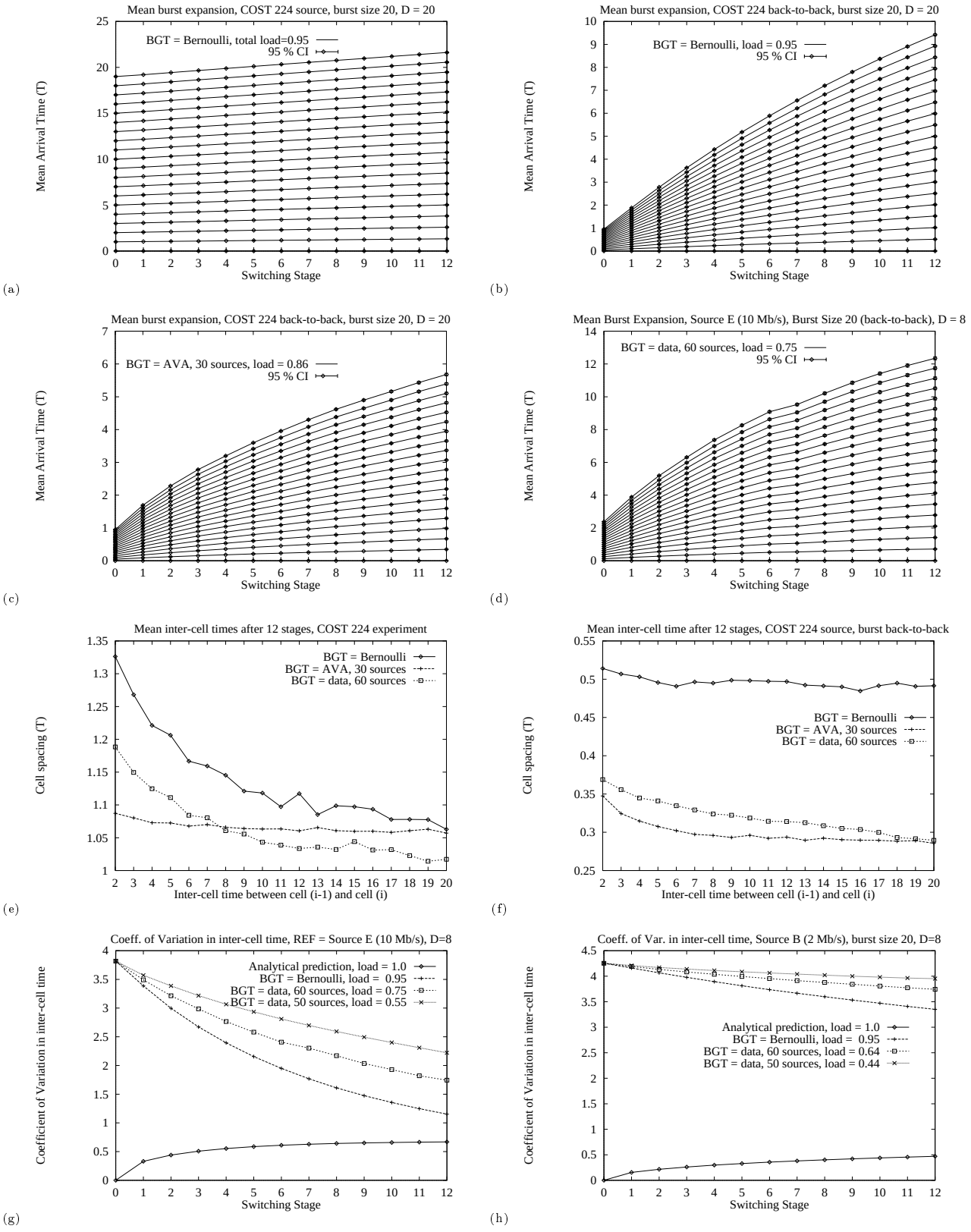


Figure 8: Burst expansion results

burst of reference traffic is composed of 20 cells transmitted within a burst time of $20T$. The bursty sources have the same mean rate as their CBR counterparts, when averaged over the burst period.

Figure 8 (b) shows the mean burst expansion for the back-to-back version of the COST 224 source, also using Bernoulli background traffic. The difference between the expansion rates of the back-to-back burst and the 1 in 20 burst used previously is enormous. The bursts expand extremely rapidly in this case, with a rate which seems to decrease as the burst expands. Figure 8 (c) shows the same source but using AVA video as the background traffic. In this case the burst expansion seems to occur at a greater rate initially, but after 12 stages the burst has expanded less than with Bernoulli traffic. From the graphs it is clear that the expansion does not proceed in a linear fashion, as conjectured in [17]. The bursts appear to expand more rapidly initially, with an upper bound on the expansion rate which is a function, once again, of the rate of the reference connection, the total load, and the background traffic type. Table 5 presents the measured expansion rates using the back-to-back COST 224 source.

Background traffic	$\rho + 1/D$	E(ΔB) (slots)
Bernoulli	0.95	14.11 (measured)
Bellcore data, 60 sources	0.68	8.53 (measured)
AVA video, 30 sources	0.86	7.88 (measured)

Table 5: Mean burst expansion, COST 224 source with back-to-back burst

Figure 8 (d) shows the burst expansion for the bursty source E. Bursts expand more quickly for the bursty source E than for bursty source B (not shown), for the same mean load. After 12 stages of switching bursts from source B expand by a factor of 9, compared to 5.6 for bursty source E, with Bernoulli traffic.

7.2 Relative Expansion within a Burst

During their simulation work to validate the burst expansion approximation, the COST 224 authors noted that the earlier cells in a burst tend to become separated more widely from the burst than later cells, as it proceeds through the network. This is because the cell arriving at the head of a burst sees a system which behaves like an $M/D/1$ queue, just before a Poisson arrival, whereas the remaining cells in the burst, in particular the last cell, see a system which behaves increasingly like an $M + D/D/1$ queue, just before a periodic arrival. The difference is quite marked, and causes increased delay for the later cells in a burst. Our results confirm this observation; the exaggerated separation of the leading cells in the burst is shown clearly in Figure 8 (e) which shows the mean inter-cell times within the bursts for COST 224 experiment, after 12 switching stages.

When the cells in a burst are transmitted back-to-back, the expansion seems to be more evenly distributed over the cells within the burst. This is shown for the back-to-back COST 224 source in Figure 8 (f). Similar observations hold for bursty sources B and E. Note when comparing the two sets of results that the inter-cell gap within a burst for the COST 224 source is T at the source, and expands to a maximum of $1.34T$ for the leading cell after 12 stages. For the COST 224 back-to-back source the initial inter-cell gap is 1 slot, or $0.05T$, expanding to a maximum of about $0.52T$ after 12 stages, for Bernoulli traffic.

7.3 Variance in Inter-cell Times

The analysis of Section 5, suggests that *irrespective* of the input traffic characteristics of a given reference source, after large number of multiplexing stages, its traffic will have the properties of a renewal process, with an inter-cell arrival time variance of $\mathrm{Var}(\mathrm{X}_\infty) = \mathrm{T}^2(\mathrm{T}-1)/(2\mathrm{T}-1)$. Thus we expect bursts of cells from a periodic bursty source to become separated by the background traffic until the inter-cell time variance is equal to $\mathrm{Var}(\mathrm{X}_\infty)$, in much the same way as we observed CBR cells to become perturbed by background traffic.

Figures 8 (g) and (h) display the measured coefficient of variation in inter-cell times for the bursty sources B and E, with Bernoulli, video and data background traffic. Each graph also shows the predicted variance using Poisson background traffic from the analytical model, assuming that the cells originate from the source in a CBR fashion. The coefficient of variation in inter-cell time of the bursty sources B and E decreases towards the bound predicted by the model as bursts proceed through the network. The results confirm that for a source with a smaller value of T (source E), the convergence to the analytical bound is more rapid than for a source with a larger value of T (source B). With bursty background traffic types and lower load the variance converges more slowly to a higher bound.

The decrease in variance can also be seen from the inter-cell departure time distributions for the same sources, in Figures 9 (a), (c), (e), and (g). Each graph has two features: a peak near the origin caused by the large number of cells which are close together within a burst (these cells are back-to-back at the source) and a peak towards the extreme of the x axis caused by the inter-cell time between the last cell in a burst and the first cell in the succeeding burst. Since it is relatively rare to have an inter-cell time greater than the inter-burst time, the expansion of the bursts takes place as cells move into the inter-burst gap from both the leading edge of the successor burst and from the tail of the previous burst. Cells tend to move more from the leading edge of the burst than from the tail (confirming our earlier results showing that leading cells are more widely separated from the burst). With Bernoulli background traffic the expansion is smooth, whilst with the other types of background traffic, however, a significant number of cells remain widely separated in time.

7.4 CDV Measurements for Bursty Sources

We conclude by studying the CDV of bursty reference sources. If we parameterise the I.371 leaky bucket as though the source were CBR with negotiated peak rate $1/T$, then the reference source will have a high CDV when it enters the network. Consider a periodic bursty CBR source whose bursts are of size b, then on entry to the network the CDV distribution will have $b-1$ peaks, at positions $i(T-1)$, for each of the i cells ($2 \leq i \leq b$) following the first cell in each burst.

What happens to the CDV as the burst progresses through the network? Our results from the previous subsections suggest that the CDV should *decrease* as the bursts expand. Our measurements confirm that the mean CDV decreases asymptotically towards a lower bound, however the same is not true of the tail of the distribution, which decreases only for Bernoulli background traffic, whereas for all other background traffic types it actually increases. This suggests that bursts can "catch up" with one another, causing high CDV.

The evolution of the CDV is pictured in Figures 9 (b), (d), (f), and (h), for bursty sources B

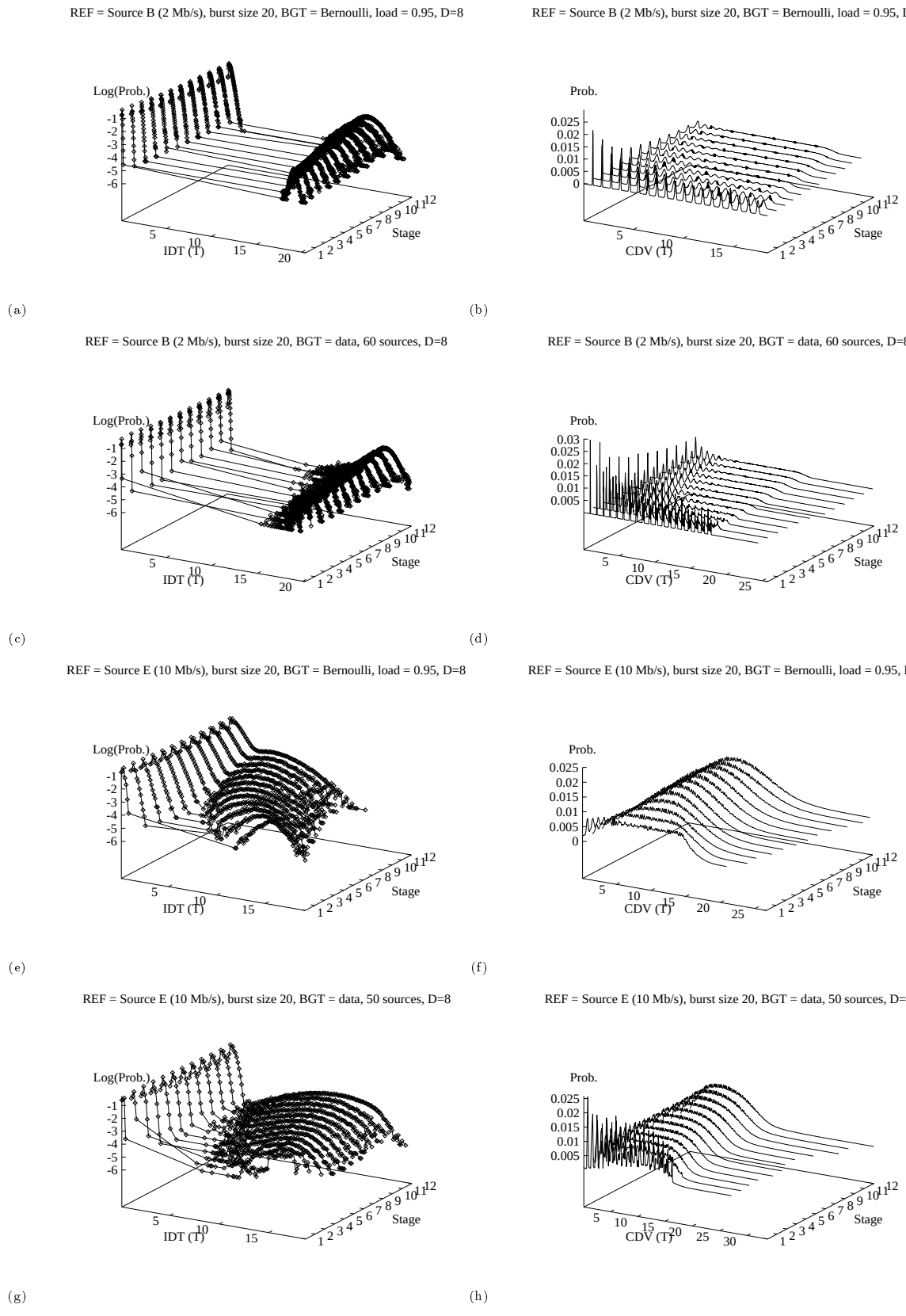


Figure 9: Inter Departure Time distributions and CDV for bursty sources B and E

and E and with two loads of data background traffic. The CDV distributions show the burst expansion clearly – the initial peaks become less visible as the reference connection cells become displaced by the background traffic. Depending on the total load and the type of the background traffic this mixing process is more or less effective at removing the burst peaks. Note that as the bursts become separated (and their mean CDV and inter-cell time variance decreases), the total total proportion of cells which experience positive CDV does not decrease – it increases but is bounded above, similar to our earlier result.

8 Non-FCFS Queueing Policies

Previous sections have studied the CDV caused by queueing at both input and output buffered switches, when the cells queued for transmission are served in First Come First Served (FCFS) order. This section briefly reviews several non-FCFS cell level scheduling policies and presents experimental results on their performance.

Two classes of non-FCFS queueing policies can be identified, namely *work conserving* and *non-work conserving*. The latter permit the server to be idle with work in the queue, with the aim of optimizing some desirable performance criterion, and have been studied only relatively recently because they are difficult to analyse theoretically.

8.1 Work Conserving Policies

All work-conserving scheduling policies use priorities of some form to distinguish between customers of different classes. Even using a simple static priority scheme, an improvement of several orders of magnitude in the loss rate of delay-sensitive traffic can be achieved with only a small performance penalty for the non-priority class if the proportion of priority traffic is low [1].

An alternative is to use priorities which vary in response to the traffic load. In classic Earliest Due Date scheduling (EDD) [10] cells are scheduled by their deadlines to the start of service. EDD can be shown to minimise the proportion of cells which miss their deadlines. Variations on EDD include the Shortest Time to Extinction [16], delay-EDD [7], and Oldest Customer First [5] policies.

Other mechanisms have been proposed which guarantee each connection a minimum service rate whilst isolating its traffic. Fair Queueing, for example, performs round robin scheduling on each of N queues, one for each connection [6], and is provably identical to Virtual Clock [24]. The visit ratio to each queue may be proportional to its requested bandwidth, and per connection queueing permits the flows to be isolated from one another.

8.2 Non-work Conserving Policies

Non-work conserving scheduling policies can provide guarantees impossible to achieve with work conserving mechanisms. "Stop-and-go" is a rate-based control which guarantees no congestion whilst guaranteeing the requested rate and the minimum and maximum end-to-end delay [8]. The jitter-EDD policy is similar to OCF, but maintains an additional variable per packet which represents the difference between its deadline and its actual service time [21]. At the next node, the packet is held for that amount of time before entering the scheduler. Thus the mechanism can bound the variance in end-to-end delay.

Hierarchical Round Robin (HRR) [11] has a time-framed structure with a hierarchy of frames. This enables the server to allocate bandwidth to a source at a requested rate. Whilst HRR guarantees delay bounds, it cannot guarantee end-to-end delay variance, because it does not reshape the traffic to its original form at every node. In addition, HRR is unable to cater for low latency, low bandwidth traffic. For further details and a comparison of different mechanisms, see [23]. Finally, a mechanism which separates the regulation of traffic and the ordering for transmission has been proposed by Zhang and Ferrari [22]. This mechanism has the advantage that it avoids the frame-based granularity of bandwidth allocation of HRR and stop-and-go, whilst providing delay, throughput and loss guarantees, and decoupling delay and bandwidth allocation.

8.3 Experiments with Delay Sensitive Traffic

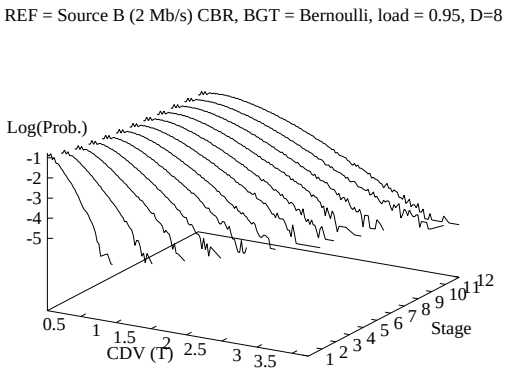
To demonstrate the difference between work conserving and non-work conserving policies, two policies were implemented: Static Priority queueing (SP) and a Rate Based scheduler (RB), equivalent to HRR. More sophisticated policies are the subject of continuing work. SP is work conserving, and RB non-work conserving. Each experiment aimed to determine the effects of one of the policies on a single delay-sensitive connection competing with the different background traffic types described in Section 4.2. As before, the connection was traced over 12 switching stages, with measurements taken after each stage.

Comparison of Policies

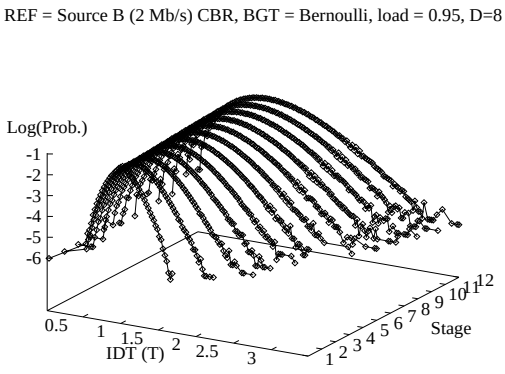
The performance of the two non-FCFS policies is best assessed in comparison with FCFS queueing. Theoretically under SP each high priority cell should be transmitted immediately and the connection properties preserved throughout. The RB policy, serving a reference connection with mean rate $1/T$, should result in a stream which is strictly periodic. Slight rate mismatches between the source and the RB scheduler will result in some variation, with the result that the inter-cell time for reference connection cells should be bounded by $2T$ [23].

Figure 10 compares the measured results for the 3 disciplines, showing the CDV and inter-cell departure time distributions after each switching stage for CBR source B (2 Mb/s), with Bernoulli background traffic and a load of 0.95. The results are startling. Superficially, RB performs best, followed by SP and FCFS. The CDV under SP is greater than expected, with a maximum of just over $2.5T$ after 12 stages, compared to $3.6T$ under FCFS. The CDV under RB is held constant but has a higher maximum value than expected, of $0.29T$. For SP the IDT distribution becomes increasingly dispersed and the initial peak (which should have been preserved throughout) is rapidly spread out.

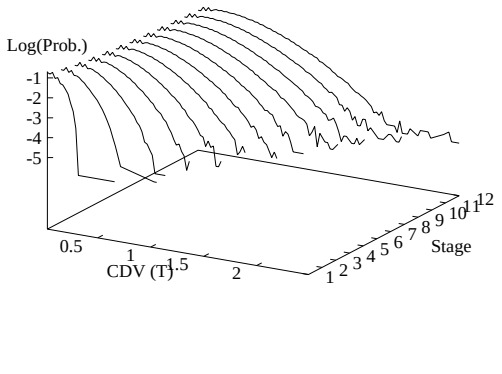
These results, together with measurements for source E (not shown), suggest that cause of the CDV is a hardware feature of the Fairisle switch. This is indeed true – the cause of the CDV under both SP and RB is a small rate adaptation buffer of maximum depth 10 cells between the output of the switching fabric and the transmission link. When this buffer contains background traffic, a priority or rate-guaranteed cell must queue in a FCFS fashion for access to the link. Fortunately under RB the effects of the FCFS queueing are corrected at each output by the strict periodicity of the scheduler, so the CDV does not degrade with additional stages. Under SP, however, there is no such correction, and the effects are cumulative. One might ask whether this is an important observation. In our experience every



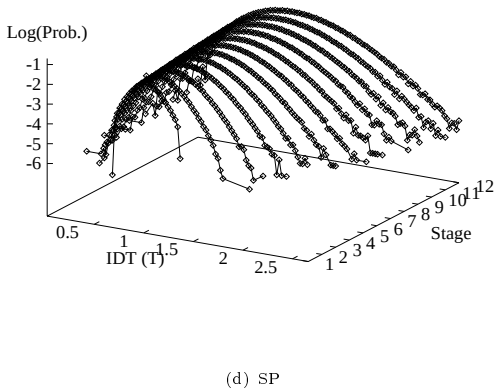
(a) FCFS



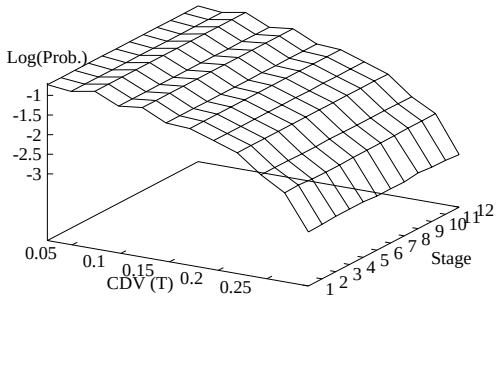
(b) FCFS



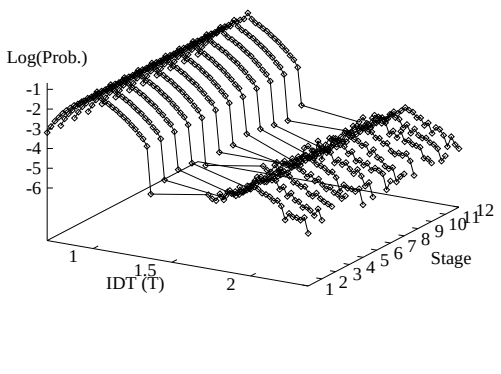
(c) SP



(d) SP



(e) RB



(f) RB

Figure 10: Comparison of CDV and inter-cell departure time distributions for source B using FCFS (a,b), SP (c,d) and RB (e,f) policies, with Bernoulli background traffic. Note the difference in CDV scales, required to make all graphs visible

piece of equipment uses such rate adaptation buffers to accommodate for the asynchronous nature of the hardware elements – we thus expect similar observations to apply to commercial ATM equipment.

When the reference connection is multiplexed with bursty background traffic the two non-FCFS policies show a clear advantage over FCFS queueing. Figure 11 shows for source E the CDV under each policy, with AVA and data background traffic. The maximum observed CDV is greatly reduced using SP, whilst under RB it is held constant (and is identical to results using Bernoulli traffic).

Finally, we observe that as expected, when a bursty reference source is used as input, the bursts are preserved by SP, but are reduced to CBR traffic by RB at the first switch. Thus RB is the most useful of the policies in dealing with bursty traffic. Space limitations preclude a full presentation of results using bursty reference sources.

9 Conclusion

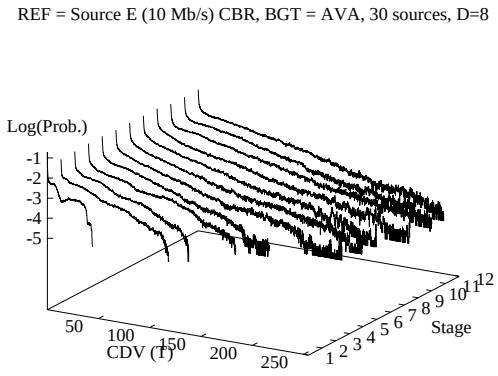
We have studied the CDV experienced by CBR connections of different rates when multiplexed with Bernoulli, data and video traffic over multiple stages of switching, comparing our results to those of analytical models and simulations. Our results confirm the broad conclusions of the analytical models, namely that the CDV is bounded and dependent on the source rate, the total load and the type of background traffic. The analytical models are conservative in their estimates and are consequently useful for network dimensioning purposes. Importantly we also conclude that Bernoulli background traffic can be a bad predictor of the performance under real traffic conditions.

We have confirmed the accuracy of two analytical approximations of burst expansion, and have shown that bursts expand less rapidly when multiplexed with bursty video and data traffic. Short term overload conditions arising from bursts of cells which are transmitted back-to-back cause the bursts to expand very rapidly. The expansion rate is determined by the source mean rate, the aggregate load and the background traffic type. Higher rate sources expand more rapidly and reach their maximum expansion rate after fewer stages of switching. Our results suggest that the CAC function need only consider CDV at the first switch – since the worst case output of the leaky bucket is a burst of cells, and bursts become dispersed, the CAC effects of CDV improve deeper into the network.

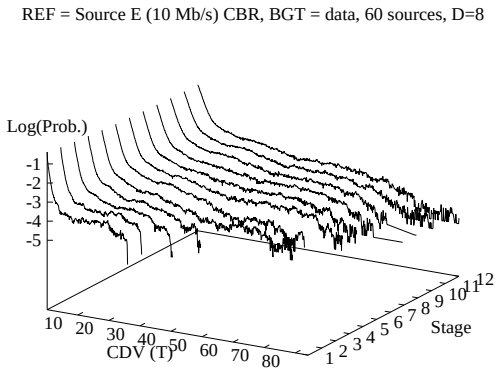
The CDV of both CBR and bursty sources is bounded. This suggests that further policing within a single network is not required. Policing at inter-network boundaries may be needed depending on the "depth" of the network.

Work conserving mechanisms can improve the performance of high priority types, but cannot prevent the occurrence of CDV. Non-work conserving policies are able to deliver guaranteed performance to different classes and can bound the CDV throughout the network. Whilst SP is attractive because of its simplicity, difficulty arises in allocating the traffic to the different priorities. Rate based schemes, particularly if positioned at network entry points, offer significant advantages, however they are more complex to implement.

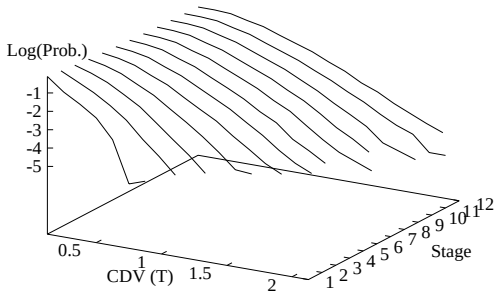
Analytical models of scheduling disciplines fail to take into account the effects of hardware properties, such as FIFOs used for rate adaptation. Whilst it might be tempting to sug-



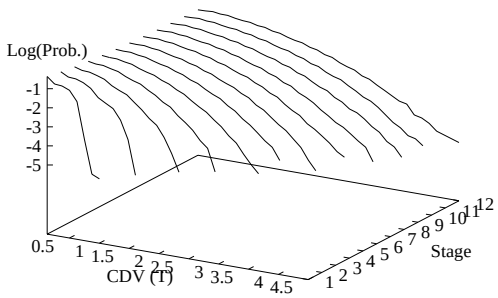
(a) FCFS



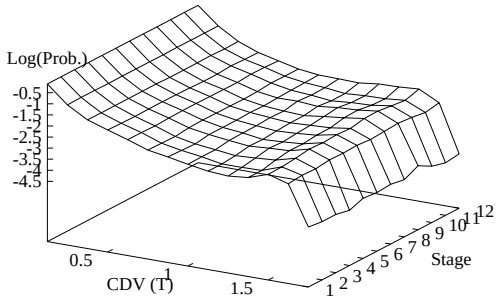
(b) FCFS



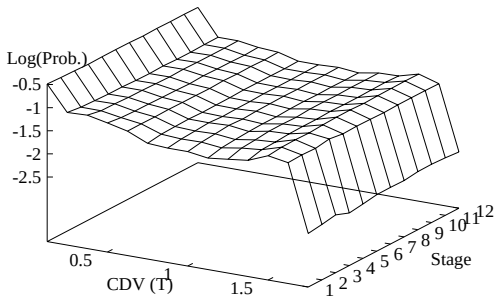
(c) SP



(d) SP



(e) RB



(f) RB

Figure 11: Comparison of CDV distributions for source E using FCFS (a,b), SP (c,d) and RB (e,f) policies, with AVA and data background traffic. Note the difference in CDV scales, required to make all graphs visible

gest that this is an experiment-specific problem, it should be clear that all ATM equipment will incorporate small buffers to permit rate adaptation. If the queueing delay these cause constitutes a significant proportion of the source rate, then they may seriously affect the performance delivered.

References

[1] J. Bae and T. Suda. Survey of traffic control schemes and protocols in ATM networks. *Proceedings IEEE*, 79(2):170–189, Feb. 1991.

[2] R. Black and S. Crosby. Experience and results from the implementation of an ATM socket family. In *Proceedings USENIX Winter Technical Conference*, San Francisco, January 1994.

[3] R. Black, I. Leslie, and D. McAuley. Experiences of building an ATM switch for the Local Area. In *Proceedings ACM SIGCOMM*, volume 24(4), September 1994.

[4] A. Brinkmann, R. Macfadyen, A. Atkinson, and E. Scharf. Validation of delay and cell loss models for ATM switches. In *Proceedings RACE EXPLOIT Traffic Workshop*, Basel, Switzerland, September 1994.

[5] T. Chen, J. Walrand, and D. Messerschmitt. Dynamic priority protocols for packet voice. *IEEE JSAC*, 7(5):632–643, June 1989.

[6] A. Demers, S. Keshav, and S. Shenker. Analysis and simulation of a fair queueing algorithm. *Proceedings ACM SIGCOMM*, pages 3–12, 1989.

[7] D. Ferrari and D. Verma. A scheme for real-time channel establishment in wide-area networks. *IEEE JSAC*, 8(3):368–379, Apr. 1989.

[8] S. Golestani. Congestion-free communication in broadband packet networks. In *Proceedings ICC*, 1990.

[9] A. Gravey and S. Blaabjerg, editors. *Cell Delay Variation in ATM Networks.* COST 242 Mid-term Seminar Interim Report, August 1994.

[10] J. Jackson. Some problems in queueing with dynamic priorities. *Naval Res. Logist. Quarterly*, 7:235–249, 1960.

[11] C. Kalmanek, H. Kanakia, and S. Keshav. Rate controlled service for very high speed networks. In *Proceedings IEEE GLOBECOM*, San Diego, California, Dec. 1990.

[12] F. Kelly and P. Key. Dimensioning playout buffers from an ATM network. In *Proceedings UK Teletraffic Congress.* IEE, 1994.

[13] W. Leland, M. Taqqu, W. Willinger, and D. Wilson. On the self-similar nature of Ethernet traffic. In *Proceedings ACM SIGCOMM*, 1993.

[14] J.-R. Louvion. First experiments on traffic and performance in the ATM pilot network. In *Proceedings RACE EXPLOIT Traffic Workshop*, Basel, Switzerland, September 1994.

[15] S. Molnár and S. Blaabjerg. The effect of multiplexing CDV affected CBR cell streams. Technical report, COST 242 TD(94), May 1994.

[16] S. Panwar, D. Towsley, and J. Wolf. Optimal scheduling policies for a class of queues with customer deadlines to the start of service. *Jnl. ACM*, 35(4):832–844, Oct. 1988.

[17] J. W. Roberts, editor. *Performance evaluation and design of multiservice networks – COST 224 Final Report.* Commission of the European Communities, Directorate-General, Telecommunications, Information Industries and Innovation, L-2920 Luxembourg, 1992.

[18] A. Roosma and W. M. ten Wolde. A new definition of cell delay variation. Technical report, Royal PTT Netherland submission to ITU-T Study Group 13, March 1994.

[19] *ATM Document Collection.* Systems Research Group, University of Cambridge Computer Laboratory, 3rd (The Blue Book) edition, March 1994. SRG Technical Note available as: http://www.cl.cam.ac.uk/Research/SRG/bluebook.html.

[20] J. L. van den Berg and J. A. C. Resing. The change of traffic characteristics in ATM networks. Technical report, COST 242 TD(92)040, 1992.

[21] D. Verma, H. Zhang, and D. Ferrari. Guaranteeing delay jitter bounds in packet switching networks. In *Proceedings Tricomm 91*, Chapel Hill, North Carolina, Apr. 1991.

[22] H. Zhang and D. Ferrari. Rate-controlled static priority queueing. In *Proceedings IEEE INFOCOM*, pages 227–234, 1993.

[23] H. Zhang and S. Keshav. Comparison of rate-based service disciplines. *Proceedings ACM SIGCOMM*, pages 113 – 121, 1991.

[24] L. Zhang. Virtual clock: A new traffic control algorithm for packet switching networks. In *Proceedings ACM SIGCOMM*, volume 20, Philadelphia, 1990. ACM.