Chapter 7

Performance for Multi-Service Traffic

The measurements presented in the previous chapter concentrated on comparing the performance of the various possible switch fabrics. We now consider how to integrate multiple services (voice, video, image, text, data, etc.) onto the switch structure. The models of the source traffic selected, and also to some extent the performance measurements taken, are necessarily simple yet they offer a first order guide to the performance of the switch under a mixed traffic load. The voice service multiplexed with data traffic is selected for a more detailed investigation as it presents a well characterised traffic source of practical interest. The results of the detailed investigation of the voice traffic model are shown to agree well with the simple model of mixed traffic but further study is required to characterise the performance of a fast packet switch for applications within the general purpose broadband ISDN.

7.1 Multi-Service Traffic Requirements

It may be argued that all communications services may be classified into two fundamental categories according to the delay requirement that they present to the network, and for lack of better terminology they will be referred to as reserved and unreserved services. A reserved service exacts an inflexible, low delay and low variance of delay requirement, whereas unreserved services are much more flexible in the range of delay that can be tolerated. Due to the delay requirement, an incoming reserved service call will only be accepted if sufficient switch bandwidth is available. A measure of the switch bandwidth allocated to reserved service calls is kept and once it reaches a maximum, determined by the delay requirement, further reserved service calls are refused. The allocation of switch bandwidth to unreserved service calls may be much less stringent.

The majority of reserved services derive from information based upon a physical property that changes rapidly with time, e.g. voice and video, and often contain a high degree of redundancy, thus permitting an appreciable packet loss before any noticeable deterioration in quality is perceived. There are some reserved services, however, that are highly sensitive to error, e.g. process control, in which the delay constraint proceeds from the requirement for a high priority service, yet such services are generally of low bandwidth. Unreserved services include the bulk of data transfer, interactive and transaction services at various priorities.

The delay constraint is not the only difference between these two basic service classifications. A reserved service requires a guaranteed bandwidth and delay performance throughout the entire duration of the connection, else the connection request must be refused. An unreserved service expects the bandwidth and delay associated with a connection to vary according to the traffic load on the network.

Three approaches to the support of these two fundamental classes of service across a fast packet switch have been proposed. The first makes no differentiation between the classes of traffic at the lowest level within the switch fabric. To guarantee the delay performance for reserved service traffic it assumes that at the access point to the network all connections are constrained to conform to the measure of network resources that each was allocated when the connection was set up. Measures such as the average and peak bandwidth and the burstiness of the traffic have been suggested [5]. This method does not share switch capacity between the various classes of traffic very efficiently and tends to limit the peak bandwidth available to bursty services.

A second approach allocates levels of priority to the different classes of traffic at the lowest level within the switch fabric. Thus high priority traffic achieves the lowest delay and variance of delay whilst the lower priority traffic shares what switch capacity remains after the higher priority traffic has been serviced. This ensures that the delay performance of the higher priority traffic is not greatly influenced by the amount of lower priority traffic within the switch. High priority (reserved service) traffic must still be allocated according to some measure of the switch capacity required but low priority traffic is not constrained to the same degree and may share the remaining available bandwidth. The priority of packets within the switch need not necessarily be allocated purely according to the class of traffic they carry. Traffic originating from some switch ports may require a higher priority than traffic from other ports. Also packets within the switch might be time stamped to give packets that have been delayed the longest in the input queues a higher priority. This would tend to reduce the variance of delay across the switch.

A third approach to the support of multi-service traffic across a fast packet switch argues that all services must be guaranteed a minimum performance at all times. The free capacity remaining within the switch fabric is measured before any connection request is granted to ensure that sufficient capacity remains to support the minimum requirements of the connection. A protocol is also implemented to ensure that the available bandwidth is distributed fairly between all switch ports and also between all classes of traffic within each switch port [54]. This approach may be the most efficient but it is difficult to apply to a large multi-path switch. For the performance measures in this chapter two classes of traffic have been introduced, reserved and unreserved, in which reserved traffic has a higher priority than unreserved traffic.

7.2 Extensions to the Switch

In order to support the two fundamental services, reserved service traffic must be given priority at all input and output ports. At the input ports, the single input queue at every port of fig. 5.2 is replaced by two queues, one for reserved service packets and one for unreserved service packets. A priority field is also added to the tag to distinguish the two classes of packet. The input port controller is modified so as to transmit unreserved service packets only when the reserved service packet queue is empty, and to postpone repeated set-up attempts of an unsuccessful unreserved service packet on the arrival of a reserved service packet. The transmission of a successful unreserved service packet is not interrupted by the arrival of a reserved service packet.

Reserved service priority must also be ensured at the output ports of the switch fabric. Two mechanisms have been investigated to implement reserved service traffic priority at the output ports and both of them are capable of a simple hardware implementation within the output port controller of the switch. The first mechanism applies to a two-plane switch structure with a simple output port controller that is only capable of handling a single packet at any one time. If a second packet arrives across the free plane while the output port controller is already busy with a packet then the set-up attempt must be rejected but the priority of the packet may be read before it is rejected. If the rejected packet is of high priority then following the completion of the packet in service the output port will only accept a high priority packet. In a two-plane switch fabric with output buffering across the switch planes an alternative algorithm is adopted. In this situation the output port controller will not accept more than one low priority packet at any one time.

7.3 Traffic Models

Two models of unreserved service traffic were used, saturation and Poisson. In the saturation model, unreserved service traffic was generated to keep each input port continuously busy while in the Poisson model, unreserved service packets were generated according to a Poisson arrival process. Both models generated traffic with a uniform random destination distribution. Three models of reserved service traffic were investigated: Poisson, talkspurt voice and TDM voice. In the Poisson model, reserved service packets were generated according to a Poisson arrival process with a uniform random distribution of packet destinations. In the talkspurt voice case, a superposition of individual voice sources was modelled, on every input port of the switch, in which the on-off characteristics of speech were used for bandwidth compression (i.e. packet voice with silence detection). Each voice source was assumed to exhibit two states, active and silent, representing the talkspurts and pauses present in conversational speech [17]. In the active state each voice source generated packets at a regular rate representing 32 kbits/sec voice coding, 256 bit packets with a further 32 bits overhead, and a 20 MHz system clock. No packets were generated in the silent state. The two states were modelled by an exponential distribution with means of 1.2 and 1.8 seconds respectively [33], and each voice source transmitted packets to a single destination which was selected at random during initialisation. The TDM voice model was simply a talkspurt model with silent periods of zero duration to represent packet voice without silence detection. A random phase relationship was assumed between all voice sources.

The measurement of delay selected for the performance of the reserved service was that of the 99^{th} percentile of the delay distribution [7]. It was assumed that packet voice traffic may withstand a 1% random packet loss, for small packet sizes [58, 60], without perceptible loss of quality. Hence, the measure of maximum delay was the delay within which 99% of all reserved service packets arrived at their destination. One consequence was that the accuracy of the maximum delay measurements was much lower than that of throughput as the tail of the delay distribution was being examined.

Delay was normalised to the packet length and all measurements were taken with a retry delay of 10% of the packet length. Where appropriate, delay results were smoothed with a least squares polynomial regression followed by interpolation with a cubic spline. Applied load and throughput per port were also normalised and reflect the average utilisation of input and output ports.

7.4 Poisson Traffic

For a 64×64 two-plane pure input buffered delta network with switching elements of degree 8, fig. 7.1 gives the throughput result with a Poisson reserved service traffic source and a saturated unreserved service traffic source on each of the switch input ports. As the reserved service traffic load is increased, so the maximum unreserved service traffic load that the switch is able to sustain falls, so as to maintain the load on the switch reasonably constant at saturation. The reserved service throughput response in the absence of any unreserved service traffic is identical to that in the presence of unreserved service sources. Fig. 7.2 gives the corresponding maximum delay curves for reserved service traffic with and without the presence of saturated unreserved service traffic. The curves are plotted as the mean of eight separate simulation runs, each run for a total of 100,000 packets at each level of reserved service traffic load, with the 95% confidence interval plotted for each point. The maximum delay for reserved service traffic in the presence of saturated unreserved service traffic is approximately 40% greater than in the absence of unreserved service traffic. This difference is due to the probability of an incident reserved service packet finding the input node already busy serving an unreserved service packet that has achieved set-up.

The same measurements were taken of a number of other switch structures but using only a single simulation run on each of the delay curves. The switch structures

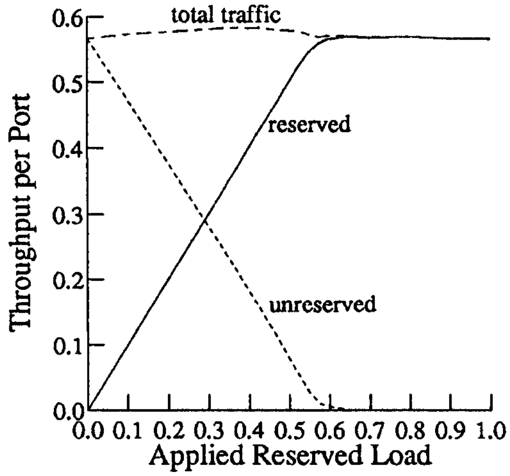

Figure 7.1: Throughput performance for the Poisson reserved service + saturated unreserved service traffic model.

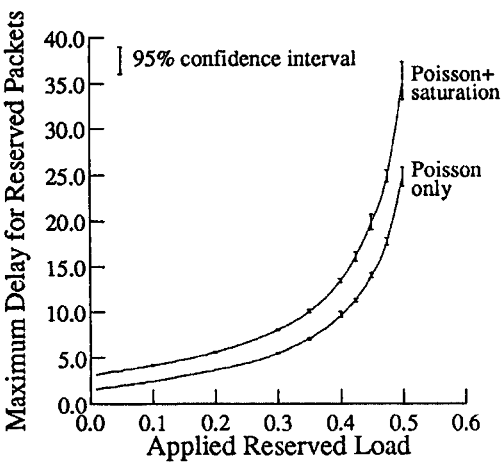

Figure 7.2: Maximum reserved service packet delay for the Poisson reserved service traffic model with and without saturated unreserved service traffic.

investigated were: the 64x64 regular two-plane delta with switching elements of degree 2 and 16; the 512 × 512 regular two-plane delta with switching elements of degree 8; the 64 × 64 sub-equipped Beneš with switching elements of degree 8 and the 64 × 64 crossbar switch. All switch structures yielded similar results scaled in proportion to the throughput at saturation of the respective switch fabric. Unless otherwise stated further measurements apply to the 64x64 pure input buffered two-plane delta network with switching elements of degree 8.

The same measurements were repeated with a non-uniform distribution of unreserved service traffic. Of the unreserved service packets, 80% were directed to 8 of the switch ports while the remaining 20% were directed randomly across all output ports. The throughput performance of the reserved service traffic was not affected but the maximum delay performance was reduced slightly. This reduction was due to the fact that the majority of the output ports of the switch were more lightly loaded with unreserved service traffic in the non-uniform distribution. It would appear that the load and distribution of the unreserved service traffic has only a slight effect on the maximum delay performance of reserved service traffic. Investigations also suggest that it is possible to operate a fast packet switch with input and output ports running at widely different mean traffic loads, as might be the case, for example, between ports connected to inter-switch trunks and those connected to local area networks. With highly non-uniform traffic distributions the total capacity of the switch is used less efficiently but the heavily utilised ports may offer a much higher throughput than for a uniform distribution.

In figs. 7.3 and 7.4 a Poisson reserved service traffic source is multiplexed with a Poisson unreserved service source at every input port of the switch. Fig. 7.3 shows the throughput performance of unreserved service traffic for several reserved service

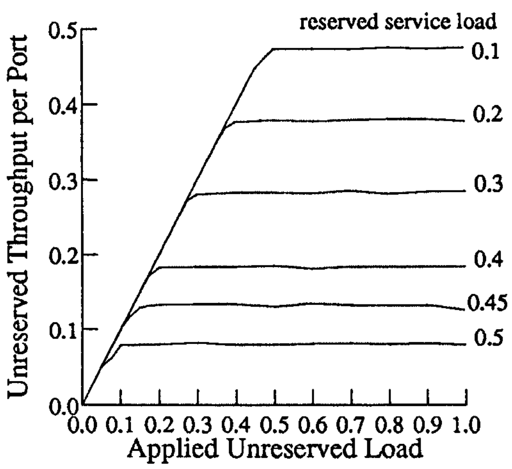

Figure 7.3: Unreserved service throughput performance for the Poisson reserved service + Poisson unreserved service traffic model.

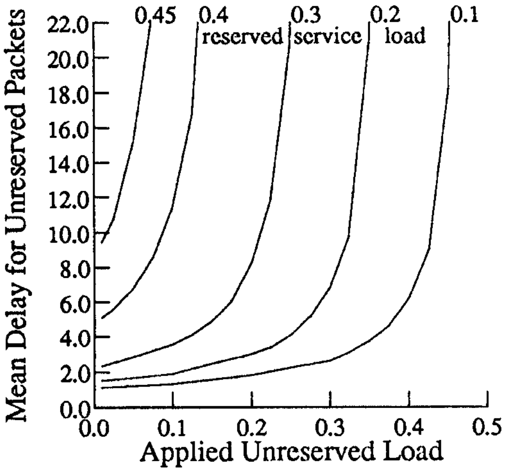

Figure 7.4: Mean unreserved service packet delay for the Poisson reserved service + Poisson unreserved service traffic model.

traffic loads. Fig. 7.4 shows the corresponding mean delay for unreserved service traffic. Both curves saturate at a level that reflects the remaining switch bandwidth available after serving the requirements of reserved service traffic. The reserved service throughput characteristic in this case is identical to that observed with a saturated unreserved service traffic source while the maximum reserved service delay is reduced in proportion to the amount that the total load on the switch falls below saturation.

To give a comparative impression of switch performance fig. 7.5(i) shows the maximum delay performance of various designs of fast packet switch of size 64×64 for Poisson traffic. Once again it may be seen that the performance of the pure output buffered switch is only slightly greater than that of the highest performance two-plane delta design. This in turn is of slightly greater performance than a two-plane Batcher-banyan (i.e. crossbar switch) as the latter is synchronous at the packet level and therefore cannot take advantage of input queue by-pass.

The performance of the two-plane delta networks of size 64x64 for Poisson reserved service traffic in the presence of saturated unreserved service traffic is presented in fig. 7.5(ii). The curves for the crossbar switch under Poisson traffic are reproduced for comparison. The maximum delay performance for reserved service traffic of the two-plane delta networks featuring output buffering is slightly impaired in the presence of saturated unreserved service traffic, particularly at low loads, due to the change in priority mechanism at the output ports. With output buffering the priority mechanism in the output port controllers will only accept a single unreserved service packet at any one time. Without output buffering an output port controller will reject unreserved service packets after detecting a failed reserved service packet set-up attempt until a reserved service packet has been serviced. The difference in performance between these two output port priority mechanisms is insufficient to be

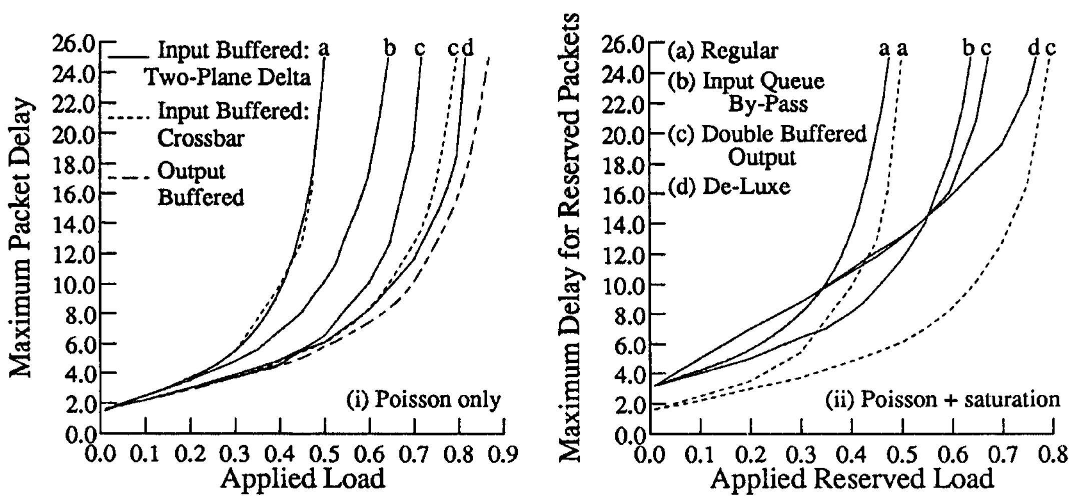

Figure 7.5: Comparison of maximum delay performance of various switch designs of size 64×64 for Poisson traffic with and without saturated unreserved service traffic.

of practical significance.

For the 64×64 regular two-plane delta network with Poisson traffic sources the queue lengths were observed to be short and to stabilise rapidly up to a traffic load of about 0.45. Beyond this traffic load the delay across the switch becomes increasingly sensitive to small changes in the applied load. This figure represents a load of 80% of the throughput at saturation of the switch fabric and provides a reasonable estimate for the upper bound of the applied reserved service traffic load for stable operation of the switch. Table 7.1 provides a comparison of the maximum delay performance of the various switch structures of size 64×64 for Poisson reserved service traffic. It presents results at the maximum load of 80% of their respective measures of throughput at saturation both in the presence and absence of saturated unreserved service traffic. For all switch structures the maximum delay in the absence of unreserved service traffic is in the region of 15 packet lengths while for structures based on the delta network in the presence of saturated unreserved service traffic it increases to about 20 packet lengths. This rule of thumb holds for all sizes and structures of switch investigated. For an implementation in CMOS operating at 50 MHz with 256 bit packets this result implies that 99% of all reserved service packets will traverse the switch within about 100 μsecs regardless of the load or distribution of the unreserved service traffic.

7.5 Talkspurt Voice

The Poisson arrival process is one of the simplest and most widely used traffic models but it is necessary to show that the performance results derived from the Poisson model bear some relation to the performance that might be expected from a more realistic model of multi-service traffic. Telephony voice traffic was chosen for a closer

Switch Design	*Throughput at Maximum Load*	*Delay at Maximum Load*	
		Poisson Only	*Poisson+ Saturation*
Two Plane Delta:			
Regular	0.453	14.3	20.3
Queue By-Pass	0.585	16.0	17.4
Double Buffer	0.639	12.2	20.0
De-Luxe	0.738	14.6	21.9
Crossbar:			
Regular	0.469	15.7	–
Double Buffer	0.718	14.2	–
Output Buffered	0.8	16.0	–

Table 7.1: Comparison of maximum delay performance of various 64×64 switch designs at maximum reserved service traffic load.

investigation as it has been widely studied, its characteristics are well known, it is fairly easily modelled to a reasonable degree of accuracy and it is of practical significance. The talkspurt voice model, described in section 7.3, was used to investigate the maximum delay performance for packet voice both with silence removed (talkspurt voice) and without the removal of silence periods (TDM voice). One problem with the modelling of talkspurt voice is the extremely long simulation runs required to simulate a reasonable number of talkspurts from each voice source. It was found that if each voice source was initialised to a random state within the talkspurt/pause cycle then the simulation would rapidly converge to a stable estimate of the maximum delay performance. Simulations at the higher loads took the longest to converge. At the highest simulated load of 80% of the throughput at saturation of the switch fabric, a simulation which ran for 50 seconds of simulated time demonstrated that a stable estimate of the maximum delay was attained after 2 seconds of simulated time. The talkspurt simulations were therefore run for 4 seconds of simulated time which was sufficient for the majority of sources to complete at least one talkspurt/pause cycle. The arrival of voice calls and the holding time of each call were not modelled. The same number of voice sources were modelled on each switch port and remained constant during each simulation run. The load on the switch was thus proportional to the number of voice sources on each switch port. With the parameters selected for the talkspurt voice model an applied load of 0.45 corresponded to 625 voice sources per switch port, and to 250 voice sources per switch port for the TDM voice model.

The maximum delay performance of the talkspurt and TDM voice models is compared to the Poisson reserved service traffic model in the absence of unreserved service traffic in fig. 7.6(a) and in the presence of saturated unreserved service traffic in fig. 7.6(b). These measurements apply to the 64×64 pure input buffered two-plane delta network with switching elements of degree 8 but measurements from other structures yield similar results. It is evident that within the region of stable operation there is no significant difference in the maximum delay across the switch

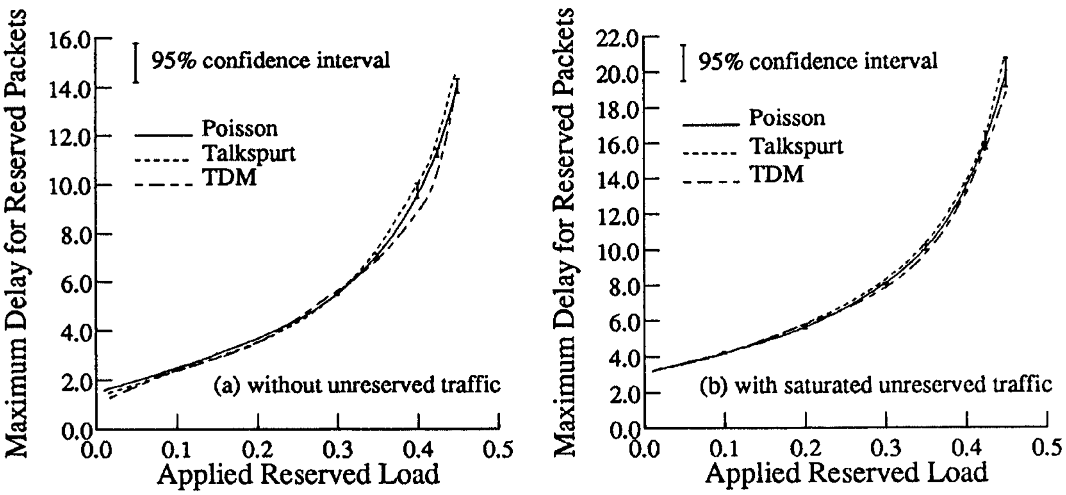

Figure 7.6: A comparison of maximum reserved service packet delay for Poisson, talkspurt and TDM voice models both with and without saturated unreserved service traffic.

for Poisson, talkspurt and TDM voice sources, either in the presence or absence of saturated unreserved service traffic. Furthermore an observation of the inter-arrival times of packets generated by the talkspurt model on a single input port reveals a very close approximation to the exponential distribution in agreement with the analysis presented in [78]. Thus the superposition of a large number of talkspurt voice sources may be modelled by a Poisson arrival process, with reasonable accuracy, for applied loads below about 80% of saturation. At loads in excess of about 80% of saturation the simulation model takes a long time to converge and analytical results suggest that its performance departs from that of the Poisson model [136, 63]. The departure from the Poisson model is to some extent dependent upon the number of sources multiplexed onto each switch port but operation at such high loads for reserved service traffic is unlikely to be required of a fast packet switch.

7.6 Packet Length

The effect of variable length unreserved service packets upon the performance of the reserved service traffic will now be examined. The results presented so far have assumed a constant packet length and all results have been normalised to become independent of the absolute packet length. In the following investigation all packets are assumed to consist of a header and an information component and the results are normalised to the value of the information component. First we consider the case in which reserved service packets and unreserved service packets are of different but constant length. The length of the unreserved service packet is expressed in terms of the reserved service packet information field, and all packets have a header of one eighth of the length of the reserved service packet information field. The throughput results

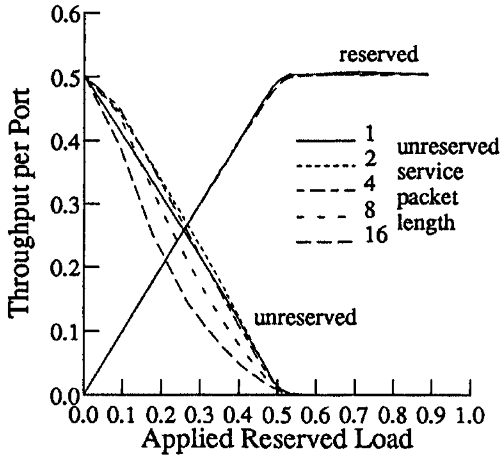

Figure 7.7: Effect of unreserved service packet length on throughput performance for the Poisson reserved service + saturated unreserved service traffic model.

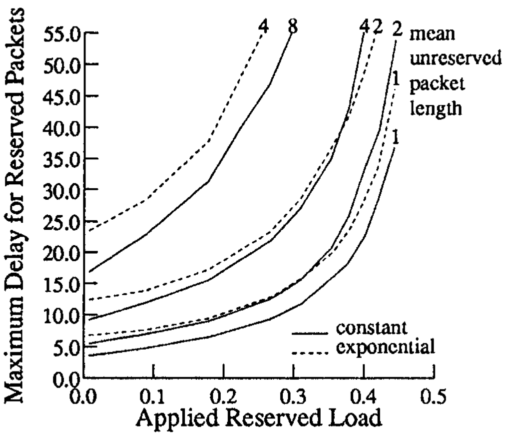

Figure 7.8: Effect of unreserved service packet length, constant and exponentially distributed, on maximum reserved service packet delay.

are presented in fig. 7.7 for the Poisson reserved service traffic model in the presence of unreserved service traffic. It may be seen that the reserved service throughput performance is not unduly affected by the unreserved service packet length. However, the unreserved service throughput at saturation, for large unreserved service packet lengths, is lower than that for small packets showing that the advantage of low packet overhead is rapidly outweighed by the superior multiplexing capability of small packet sizes.

If this result applied to multiplexed flows in general it would indeed be very interesting, however, it is likely to be the result of head of the line blocking in an input buffered fast packet switch with non-preemptive reserved service traffic priority. As the packet length of unreserved service traffic increases, so the probability of head of the line blocking in the reserved service input queues is increased. This in turn reduces the switch capacity available to unreserved service traffic. An examination of the case in which the length of the information component of all unreserved service packets was given by an exponential distribution revealed similar results with a further reduction in the unreserved service throughput performance of between 10% to 20%, due to the variability in packet length.

The effect of the unreserved service packet length upon the maximum reserved service packet delay performance is given in fig. 7.8. As expected, a variable length unreserved service packet exerts a greater impairment of performance than one of constant length, and the shorter the mean packet length the less the reserved service packet delay performance is affected. Hence, conventional sizes of data packet must clearly be broken down into short packets for multiplexing with delay sensitive traffic but this may not be necessary for a ‘data-only’ environment.

The throughput and maximum delay performance was also considered of Poisson traffic in the absence of unreserved service traffic in which all packet lengths followed a uniform random distribution of ±10% about the mean value. No drop in performance was detected with respect to that of constant length packets. For the case in which all packet lengths followed an exponential distribution about the mean, a drop in the throughput performance of about 12% was measured with a corresponding impairment of the delay performance. Thus the switch is insensitive to the variation in packet length that might be introduced by a line code employing 'bit-stuffing' but larger variations will cause a reduction in performance.

7.7 Buffer Overflow

In the performance measurements presented so far in this chapter no constraint has been imposed upon the length of the input queues. In practice an upper limit of 100 packets was imposed upon all queues but under normal operating conditions this limit was never approached. It is interesting to consider the maximum length of input queues required to maintain a given probability of buffer overflow.

An approximate analysis for slotted traffic is presented in [71] in which the buffer overflow probability is given by the expression:

$$\frac{p(2-p)}{2(1-p)}\left[\frac{p^2}{2(1-p)^2}\right]^B$$

The applied load is given by p which represents the probability of an input timeslot containing a packet while B gives the number of packet buffers in each input queue. Using a simulation model to measure the packet loss probability a minimum buffer overflow of 100 packets or so is required in order to yield a measurement of reasonable accuracy. This makes the measurement of packet loss probabilities better than 10^{-4} difficult using a simulation model without excessively long simulation runs [5]. A linear regression of the approximate simulation results for the $\log_{10}$ of the buffer overflow probability is given in fig. 7.9 for buffer lengths of 8, 12, 16, and 20. These results derive from an asynchronous crossbar switch fabric under slotted traffic with a retry delay of 10% of the packet length. The curves are extrapolated to a packet loss probability of 10^{-6} and compared to the analytical result for a synchronous crossbar switch fabric. The results differ due to the difference in throughput performance between the synchronous and asynchronous crossbar switch models and also because both results are only approximations. Further investigation suggests that at a load of 80% of the throughput at saturation of the switch fabric the single plane regular delta structure has a slightly poorer packet loss performance than the crossbar switch and the two-plane delta network with output buffering a slightly better performance.

To gain a more accurate insight into the packet loss probability of the various switch designs a simulation model would be required designed specifically for that task. These results, however, indicate that at a load of 80% of saturation, with input buffers of length 20 packets, all switches offer a packet loss probability better than

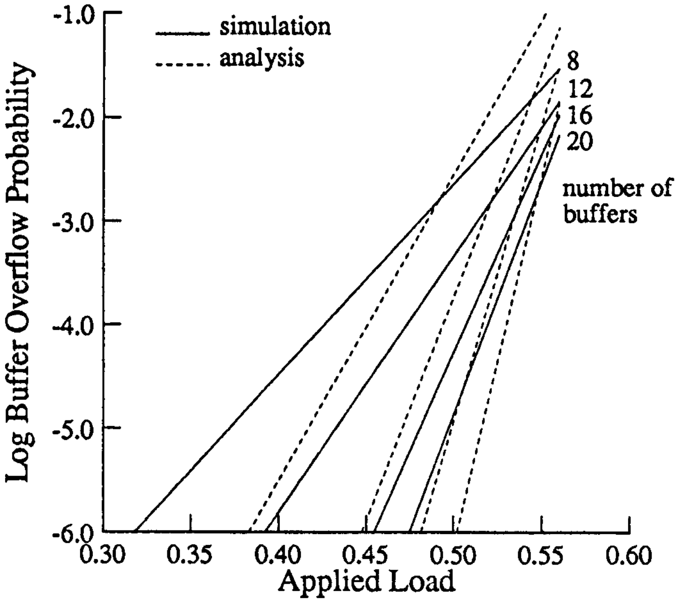

Figure 7.9: Buffer overflow probability for the input buffered crossbar switch.

10^{-6} for slotted traffic. In the case of two-plane switches with output buffering the performance appears to be significantly better than 10^{-6}.

7.8 Discussion

The major result of the performance investigation of the fast packet switch for multi-service traffic is that if a minimum of two classes of traffic are defined a statistical guarantee can be made concerning the delay performance of the higher priority class of traffic. For all switch designs investigated 99% of all reserved service packets will traverse the switch within a delay of less than 20 packet lengths provided that the reserved service traffic load does not exceed 80% of the throughput at saturation of the switch fabric. This guarantee holds regardless of the traffic load or distribution of the unreserved service.

The reserved service is so named because before an incoming call is accepted the switch checks to see that the required bandwidth is available before allocating it to the new call. If insufficient bandwidth is available the call is refused. To accomplish this the switch may be able to make peak and average load measurements on the input and output ports of the switch required by the new call. Alternatively it might maintain a sum of the bandwidth already allocated on all input and output ports. The new call supplies an estimate of the bandwidth resources required and provided that both input and output ports may accommodate this load the new call is accepted. For this bandwidth reservation mechanism to work each call must be monitored by a policing mechanism to ensure that it does not exceed the bandwidth resources that it was allocated when accepted. One possible method of implementing this control mechanism at the entry point to a network of fast packet switches is discussed in [5] which uses three parameters to characterise the bandwidth requirements of a call: peak bandwidth, average bandwidth and burstiness. From these parameters a method of estimating the effective bandwidth required by a call is discussed and a simple hardware mechanism is given to ensure that the call does not exceed the agreed parameters.

Some sources of reserved service traffic are easily characterised in terms of their traffic requirements. Voice telephony is an obvious example in which the traffic characteristics depend upon the coding scheme employed, but for a large enough multiplex of sources, reliable statistical assumptions may be made. Video sources are not so easily characterised and their traffic requirements depend heavily upon the coding employed and the picture content. Video represents a class of traffic that may be both very bursty and also exhibit a low delay requirement. One solution for handling such traffic is to extend the priority scheme to more than two levels and to allocate video traffic a priority below voice traffic but above delay insensitive traffic. An estimate of the average bandwidth requirement of a video call must be made by the bandwith reservation mechanism and the entire class of video traffic must be subjected to some form of policing mechanism to ensure that it does not exceed its bandwidth allocation [2]. The parameters of the policing mechanism might be varied according to the load of lower priority traffic in the switch. This would permit greater flexibility when the switch was more lightly loaded. Another possibility for handling video sources is to use variable bit rate coding with interaction between each source and the network to vary the source traffic characteristics during a call according to the bandwidth available across the network.

It must also be recognised that the switch is a statistical switch and that the measure of delay performance for reserved service traffic is an average measure taken over all ports of the switch for the duration of the measurement. There will be some shorter periods of time in which the delay performance is worse than the average. Also the delay performance is averaged over all calls on each switch port thus some calls may receive a poorer performance than others during the lifetime of the connection. In order to reduce this effect it may be necessary to select a maximum load for reserved service traffic below 80% of the throughput at saturation depending upon the delay sensitivity of the source traffic and the sensitivity of the switch to short duration overloads.

When bandwidth is allocated to the reserved service it is not removed from the pool of bandwidth available to both reserved and unreserved service calls as would be the case in a circuit switch. All of the switch bandwidth is shared between all traffic, with reserved service packets receiving the higher priority. If the unreserved service traffic is also to be assured of a worst case delay performance, the unreserved service calls will also need to be monitored and bandwidth allocated but probably in a less stringent manner than for the reserved service.

Although only two classes of traffic have been investigated it is clear that the technique may be extended to a larger number of classes of traffic. Internal network signalling traffic, for example, may require the highest priority while the unreserved service may be divided into interactive services which have some delay requirement and bulk transfer services that are much less sensitive to delay.

7.9 Summary

An extension to the design of the switch has been proposed to support two fundamental classes of traffic. Reserved service traffic receives a higher priority in the switch fabric and handles classes of traffic that are sensitive to delay whilst the unreserved service caters for traffic that is less delay sensitive. Simulation results indicate that for a Poisson reserved service traffic loading of up to 80% of the throughput at saturation of the switch fabric, the upper bound on delay for 99% of all incident reserved service packets is in the region of 20 packet lengths. Further, unreserved service traffic may be multiplexed with reserved service traffic, at every input port of the switch, so as to operate the switch continuously at saturation, without affecting the bounded delay performance of the reserved service. This result has been shown to hold for a wide range of switch structures and switch fabric design parameters. Also the reserved service throughput and delay performance appears insensitive to the arrival distribution and to the destination distribution of unreserved service traffic.

A closer investigation of the delay performance of the switch for voice traffic modelled as a superposition of individual packet voice sources on each switch port, both with and without silence detection, reveals no significant departure from the delay performance of the Poisson model. This traffic model was observed to give a packet arrival distribution closely approximating that of a Poisson source.

For delay sensitive, reserved service traffic performance, the packet length for both reserved and unreserved service traffic should be kept short and constant. No performance impairment is introduced by a ±10% variation in packet length but an exponential distribution of packet lengths causes a loss in throughput performance of the order of 12% for a 64×64 regular two-plane delta network. For a single service implementation, moderately insensitive to delay, variable length packets of any reasonable maximum length may be supported.

A cursory inspection of the buffer overflow probability suggests that an input buffer length of 20 packets is sufficient to offer a packet loss probability of less than 10^{-6} for slotted traffic at a traffic load of 80% of the throughput at saturation for all switch designs.