Input Queued Switches for Variable Length Packets: Analysis for Poisson and Self Similar Traffic

D. Manjunath

Department of Electrical Engineering, Indian Institute of Technology, Bombay, Powai, Mumbai 400 076 India

Biplab Sikdar

Department of ECSE, Rensselaer Polytechnic Institute, Troy, NY 12180 USA

Abstract

We consider non blocking, variable length packet switches where packet lengths and interarrival times have continuous distributions as is applicable in IP networks. A general throughput-delay model for Poisson and self similar packet arrivals of exponential lengths to a single stage $M \times N$ switch with infinite and finite buffers is obtained. Analytical results are compared against simulation results on traces statistically similar to Bellcore traces. Tradeoffs between fixed packet length VOQ switches and variable length FIFO-CIOQ switches with speedup and parallelism (multiple switching planes) are also studied. Analysis shows that a parallelism of 4 achieves 99.9% throughput. We also analyze the effect of traffic asymmetries and hotspots.

Key words: Variable Length Packet Switches, IP Switching, Self Similar Traffic.

1 Introduction

With increasing throughput and port density requirements in the Internet, much of the switching will have to be done in hardware, possibly using space switches, preferably those with the nonblocking property. It is now well known that output queued (OQ) switches can provide 100% throughput and also provide arbitrary QoS provisions efficiently. However, they are deemed infeasible to implement at high speeds and high port densities where they will be limited by the switch and memory speed requirements. Thus there is considerable interest in architecture and performance of input queued (IQ) switches.

Preprint submitted to Elsevier Science *5 September 2001*

Much of the switching literature, be it those that analyze the performance, propose new designs and architectures or explore QoS enabling properties of these switches assume fixed length packets. For example, for fixed packet lengths with time slotted operation of the switch, (typically slot size will be equal to packet lengths), discrete time analyses under various assumptions are available. Patel [29] shows that in such switches with no input or output buffers under Bernoulli packet arrivals the maximum throughput is $1 - 1/e \approx 0.63$. Karol et al [15] show that the saturation throughput of an input queued switch with infinite input buffers is $2 - \sqrt{2} \approx 0.586$. Li [19] shows that when the arrivals are correlated, the saturation throughput of the input queued switch is 0.5. In an unpublished paper, Kumar and Jacob [17] show that the saturation throughput is indeed the capacity of the switch.

The throughput of an input queued switch can be increased by having multiple queues at each input which allows multiple packets, possibly with different destinations to compete from each input. A virtual output queued (VOQ) switch takes this idea to the extreme – each input maintains a separate queue for each output. A central scheduler selects the packets to be switched in each slot such that at most one packet from each input and at most one packet for each output is selected in a slot. McKeown *et al* [22] show that if the packet arrivals to the switch are *iid*, 100% throughput can be achieved in a VOQ switch by using a maximum weight matching algorithm to select the packets to be switched in each slot. Weights could be either the length of the virtual output queue or the waiting time of the queue at the head of the virtual output queue. More practical scheduling algorithms that ensure that each of the N^2 queues is equally likely to be selected in a slot and that none of them starve have been reported in [2,21,23–25,5,6]. Of these Parallel Iterative Matching (PIM), [2], iSLIP [25] and dual round robin [5] algorithms have been implemented in the AN2 [2], Tiny Tera [26] and Saturn [5] switches respectively.

Increasing interest in providing QoS in the Internet, has led to a corresponding interest in the QoS provisioning properties of input queued switches. Recent results show that a VOQ switch with a speedup of two and a centralized scheduler can match the output sequence of an OQ switch with a QoS scheduler on each port [7,16,27]. Also, Dai and Prabhakar [8] generalize the result of [22] to show that with a speedup of two, any maximal matching algorithm will achieve 100% throughput under any packet arrival process. Summarizing this discussion, we see that to increase the throughput for arbitrary arrival processes in the VOQ switch and also to be able to provide QoS, it is essential that the VOQ switch operate with a speedup of at least two, i.e., we need to use a combined input/output queued (CIOQ) switch. Thus in a CIOQ switch with virtual output queues, in comparison with an output queued switch, the implementation complexity, is shifted from the switch to the centralized scheduler (for QoS and also for maximal matching) which needs to have information about the occupancy (and QoS requirements) of the N^2 queues, i.e., a VOQ switch may not be much less complex than a VOQ after all. (Note that a 40Gbps, scalable to 160Gbps, 64×64 OQ switch is reported in [32].)

With IP dominating the Internet, packet arrival times at IP packet switches and packet lengths will be drawn from continuous nonnegative distributions. This is in contrast with ATM switches that have a time slotted operation, fixed packet lengths and packet arrivals at slot boundaries. Also, note that the results discussed above are for fixed length packets with time slotted operation of the switch. To switch variable length packets, the obvious thing to do is to break them up into fixed length units, call them cells, switch the cells and then reassemble the cells at the output. Such a switch should have a speedup greater to handle the increased load resulting from the padding of packets that don't fill an integral number of cells. There is also the additional requirement of circuits to disassemble and reassemble packets. This leads us to believe that for variable length packet switches with non slotted arrivals, a FIFO-CIOQ switch with speedup will be architecturally simpler than the VOQ-CIOQ switch and will achieve low latencies in the input queue, which in turn will enable it to use output port QoS schedulers. A further alternative to increase the throughput of FIFO-CIOQ switches would be to have parallel switching planes, such that more than one packet is switched to an output queue *simultaneously* while from each input at most one packet is selected. This latter feature of the switch architecture will be called parallelism. In this paper we develop analytical models for the throughput, delay and loss performance of a variable length FIFO-CIOQ switch with speedup and parallelism.

There is surprisingly little study of non time-slotted variable length packet switch performance and architecture. To the best of our knowledge, the only architecture that does not breakup variable length packets into smaller cells is reported by Yoshigoe and Christensen [35]. Likewise, the only performance model for variable length packet switches is by Fuhrman [12]. He considers a $M \times N$ nonblocking input queued variable length packet switch. For *iid* Poisson arrivals processes at each node and uniform routing it is shown that the saturation throughput per port is $M/(M+N+1)$. An approximate mean delay analysis is also presented. The first part of this paper can be considered to be a generalization of the results of [12] where we present a throughput delay analysis for an $M \times N$ input queued switch with arbitrary Poisson arrivals at each input, exponential packet lengths, arbitrary output line rates and arbitrary routing probabilities.

It is now well known that packet arrival processes in the wide area exhibit *long range dependence* (LRD) whose effects are not captured by Poisson models. The correlations and burstiness over many time scales of LRD packet arrivals impact the queuing performance in a manner considerably different from those with Poisson arrivals because of extended periods of large queue buildups.

Thus the interaction of the LRD packet arrivals and HOL blocking in a switch can lead to very bad queuing behavior. In view of the extreme queuing behavior expected, a deeper understanding of the switch behavior becomes necessary because the switch is the critical component in providing various quality of service guarantees in the multiservice Internet of the future. Variable length packet switches have not been analyzed for non Poisson packet arrivals. In this paper we also consider Poisson and LRD packet arrivals and develop throughput, delay and loss models for input queued, variable length packet switches with LRD packet arrivals.

This paper is organized as follows. In Section 2 we present the delay throughput analysis for a $M \times N$ switch with Poisson arrivals and exponentially distributed packet lengths. We consider cases of both infinite and finite input buffers and present the loss analysis for the finite input buffer case. In Section 3 we present the delay throughput analysis for a $M \times N$ switch with self-similar interarrival times at each of the inputs. Once again, we consider both infinite and finite buffer switches. In Section 4 we discuss the comparative merits of the techniques used to increase the throughput in input queued switches and present some results from our analysis for combined input output queued switches. In Section 5 we consider the effects of output hotspots and analyze the switch behavior in the presence of such hotspots. In Section 6 for a given arrival rate, we consider the effect of the various parameters of the self similar process on the performance of the switch. Finally, in Section 7 we present a discussion on the results and concluding remarks.

2 $M \times N$ Switch with Poisson Arrivals, Exponential Packet Lengths

We first consider a single stage unslotted, internally nonblocking $M \times N$ input queued packet switch. Packet arrivals to input port i form a Poisson process of rate λ_i and choose a destination j with probability p_{ij}. The line rate on output port j is μ_j and there are no buffers at the output. Input packets are served according to FIFO. When a packet moves to the head of its queue, if its destination is busy, the packet will wait at the head of the input queue till the destination output port is free and chooses to evacuate the packet. When an output port finishes service, of the packets that are waiting at the head of the queues of the inputs, the packet that was blocked first is served first. Service in random order, round robin or processor sharing disciplines can also be analyzed using the method developed here but we do not investigate them. From above, the arrival rate to output port j, Λ_j, and its utilization, η_j, are

$$\Lambda_j = \sum_{i=1}^{M} \lambda_i p_{ij} \qquad \eta_j = \frac{\Lambda_j}{\mu_j} \tag{1}$$

The sojourn time of an input packet has two components - waiting time in the input queue till it moves to the head of the line (HOL) and the time spent at the HOL of the input queue till the HOL packets from other input queues that were blocked earlier finish their service and the packet is evacuated. The time spent at the HOL of the input queue corresponds to the "service time" in the input queue. This service time, once again, has two components - a blocking delay, the time until the output starts evacuating it, and the actual service time, the time taken to evacuate the packet by the destination port. Figure 1 shows these times in detail. Since the arrivals to the input queue are Poisson, each input queue can be seen to be a M/G/1 queue with service time distribution given by the time spent by a packet at its HOL. To analyze the queuing behavior the distribution of the time spent at the HOL of the queue needs to be obtained and this is derived below. In this derivation, we use techniques similar to the analysis of queueing networks with blocking [31].

Consider output port j. It has room for only the packet that is being evacuated (served). However, the HOL positions at the M input queues can contain a packet meant for output j which are waiting for the port to become free. These packets form a virtual queue for output j and are served FCFS. Thus the virtual queue of any output has at most M buffers. The time taken by the output port to evacuate a packet from the HOL of the inputs is exponentially distributed with mean $1/\mu_j$. If we approximate the arrival process to the virtual queue by a Poisson process of throughput Λ_j, then output queue j can be modeled as a M/M/1/M queue. Since the queue has finite buffers, the throughput is not equal to the arrival rate. The throughput of output port j should be Λ_j. Therefore the "arrival rate" corresponding to this throughput, let us call this the effective arrival rate and denote it by Λ'_j, will be obtained by solving for Λ'_j in the equation

$$\Lambda_j = \Lambda'_j \left[1 - \frac{1-{\eta'}_j}{1-{\eta'}_j^{M+1}}{\eta'}_j^{M}\right] = \Lambda'_j \frac{1-{\eta'}_j^{M}}{1-{\eta'}_j^{M+1}} \tag{2}$$

where $\eta'_j = \Lambda'_j/\mu_j$. The term in the square brackets in the first equality corresponds to the probability that an arriving packet into an M/M/1/M queue is not blocked.

The probability that there are k packets in the virtual queue of output port j, $\theta_j(k)$, is given by

$$\theta_j(k) = \frac{(1-\eta'_j)(\eta'_j)^k}{1-(\eta'_j)^{M+1}} \qquad \text{for } k = 0 \cdots M \tag{3}$$

Packet arrivals to the head of an input queue are approximated to form a Poisson process. Thus the probability that it will see k packets ahead of it

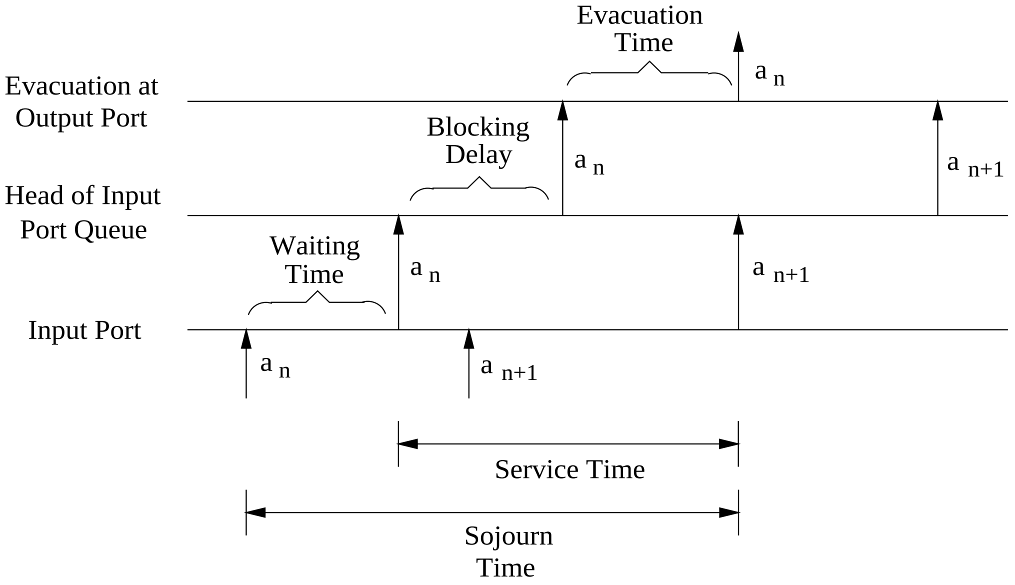


Fig. 1. Time diagram for the sojourn time in the switch. a_n represents the n^{th} arrival to the input queue and all times shown in this figure correspond to this packet.

in the virtual queue of the output will be $\theta_j(k)$. However a packet moving to the head of an input queue can see only $0, 1, \cdots M-1$ and will never see M packets ahead of it. Therefore the probability that a packet arriving to the head of an input queue wanting to go to output j sees k packets ahead of it, $\pi_j(k)$, will be

$$\pi_j(k) = \left[\frac{\theta_j(k)}{1-\theta_j(M)}\right] = \left[\frac{(1-\eta'_j)(\eta'_j)^k}{1-(\eta'_j)^M}\right] \qquad \text{for } k = 0, 1, \cdots M-1 \tag{4}$$

In the virtual queue of output port j if there are k packets ahead of it, the packet has to wait for the evacuation of these packets before it can begin its service and its waiting time is a k stage Erlangian distribution (sum of the k independent, exponentially distributed evacuation times). In addition to the blocking delay there is the evacuation time that has an exponential distribution of mean $1/\mu_j$. Thus the conditional (conditioned on the packet wanting to go to output port j) sojourn time of a packet at the HOL of the input queue has a phase type distribution like that shown in Figure 2. The Laplace-Stieltjes Transform (LST) of the unconditional distribution of the sojourn time at the head of input i, $X_i(s)$, can be seen to be

$$X_i(s) = \sum_{j=1}^{N} p_{ij} \left[\sum_{k=1}^{M-1} \pi_i(k) \left(\frac{\mu_j}{\mu_j+s}\right)^k\right] \left[\frac{\mu_j}{\mu_j+s}\right] \tag{5}$$

Here the term in the first square brackets corresponds to the blocking delay and that in the second corresponds to the evacuation time given that the

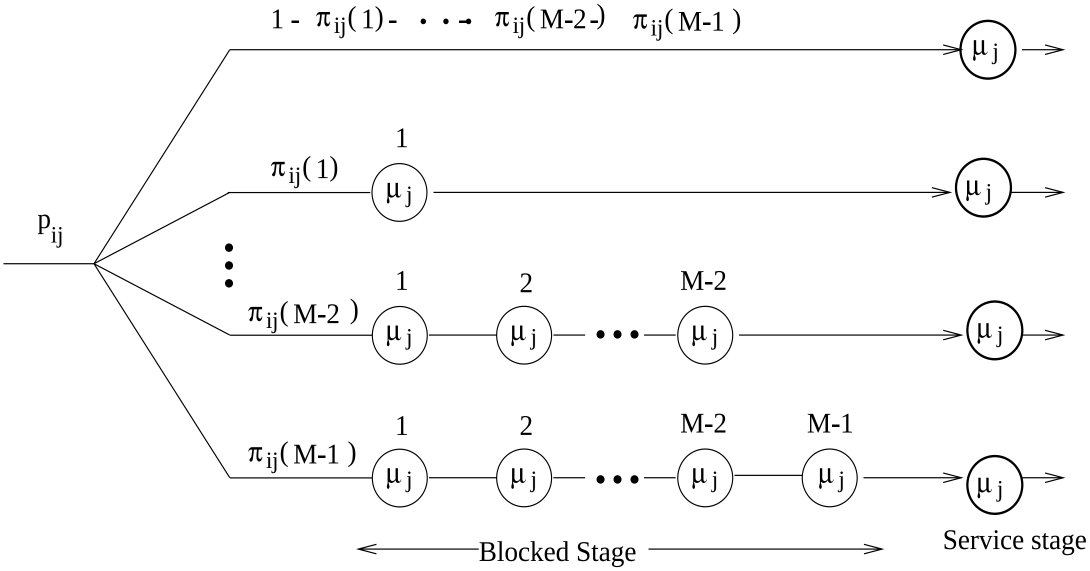


Fig. 2. Phase-type distribution for sojourn time in virtual queue of output j when a packet reaches the HOL of input i. The blocking delay is a k-stage Erlangian with probability $\pi_{ij}(k)$, the probability that there are k packets in the virtual queue of output port j ahead of this packet. There is an additional service stage corresponding to the evacuation of the packet from the input queue by the output port.

packet wants to go to output j. The first three moments of the blocking delay at input queue i, $\overline{B_i}$, $\overline{B_i^2}$ and $\overline{B_i^3}$ respectively, are

$$\begin{aligned}
\overline{B_i} &= \sum_{j=1}^{N} p_{ij} \sum_{k=1}^{M-1} \pi_j(k) \frac{k}{\mu_j} \\
\overline{B_i^2} &= \sum_{j=1}^{N} p_{ij} \sum_{k=1}^{M-1} \pi_j(k) \frac{k(k+1)}{\mu_j^2} \\
\overline{B_i^3} &= \sum_{j=1}^{N} p_{ij} \sum_{k=1}^{M-1} \pi_j(k) \frac{k(k+1)(k+2)}{\mu_j^3}
\end{aligned} \tag{6}$$

Likewise, the first three moments of the service time for the input queue, $\overline{X_i}$, $\overline{X_i^2}$, $\overline{X_i^3}$ respectively, are

$$\begin{aligned}
\overline{X_i} &= \overline{B_i} + \sum_{j=1}^{N} \frac{p_{ij}}{\mu_j} \\
\overline{X_i^2} &= \overline{B_i^2} + 2\overline{B_i} \sum_{j=1}^{N} \frac{p_{ij}}{\mu_j} + 2 \sum_{j=1}^{N} \frac{p_{ij}}{\mu_j^2} \\
\overline{X_i^3} &= \overline{B_i^3} + 3\overline{B_i^2} \sum_{j=1}^{N} \frac{p_{ij}}{\mu_j} + 6\overline{B_i} \sum_{j=1}^{N} \frac{p_{ij}}{\mu_j^2} + 6 \sum_{j=1}^{N} \frac{p_{ij}}{\mu_j^3}
\end{aligned} \tag{7}$$

and the mean sojourn time in the switch for an input packet to port i, $\overline{D_i}$, is (from the Pollaczek-Khinchin formula)

$$\overline{D_i} = \frac{\lambda_i \overline{X_i^2}}{2(1-\lambda_i \overline{X_i})} + \overline{B_i} + \sum_{j=1}^{N} \frac{p_{ij}}{\mu_j} \tag{8}$$

The maximum arrival rate that input port i can support is obtained by solving for λ_i in $\lambda_i \overline{X_i} = 1.0$. The second central moment of sojourn time in the switch for an input packet to port i, $\overline{D_i^2}$, is easily calculated from the moments of the waiting time, the blocking time and the evacuation times and can be shown to be

$$\overline{D_i^2} = \frac{\lambda_i \overline{X_i^3}}{3(1-\lambda_i \overline{X_i})} + \frac{\lambda_i^2 \overline{X_i^2}^2}{2(1-\lambda_i \overline{X_i})^2} + \frac{\lambda_i \overline{X_i^2}}{(1-\lambda_i \overline{X_i})} \left(\overline{B_i} + \sum_{j=1}^{N} \frac{p_{ij}}{\mu_j} \right) + \overline{B_i^2} + 2\sum_{j=1}^{N} \frac{p_{ij}}{\mu_j^2} + 2\overline{B_i} \sum_{j=1}^{N} \frac{p_{ij}}{\mu_j} \tag{9}$$

Consider the special case of an $N \times N$ switch with $p_{ij} = 1/N$ for all i, j; $\lambda_i = \lambda$ for all i and $\mu_j = 1.0$ for all j. Figure 3 shows the total delay and the blocking delay for various values of N from analytical and simulation models as a function of λ. Note that the difference between the analytical and simulation models improves for both the total and the blocking delay as the switch size increases. For verifying the accuracy of the analytic models, we developed a simulator working in continuous time with packets arriving at the input ports according to a Poisson with user specified rates. The packet lengths were generated according to an exponential distribution. For the results in Figure 3, the simulator had infinite buffers at the input ports. This simulator is the basis of the simulation results presented in this paper and was modified appropriately for generating the results for the finite buffer, speedup and self-similar arrivals. The simulation is stopped when the statistics that we are interested in converges to within two decimal places. All the simulations for this paper were carried out in this manner.

It is easily seen that our delay model is exact for $N \to \infty$. As $N \to \infty$, the virtual M/M/1/N queue of the outputs becomes an M/M/1 queue with arrival rate λ and service rate 1.0. As $N \to \infty$, the arrival process to the input queue is Poisson with rate λ and it in turn is an M/G/1 queue with service time equal to the sojourn time in an M/M/1 queue with arrival rate λ and service rate 1.0. Thus for the input queue to be stable, λ should be less than the reciprocal of the sojourn time of an M/M/1 queue with arrival rate λ and service rate 1.0. This yields the condition, $\lambda \leq 1 - \lambda$ or $\lambda < 0.5$ for stable queues at the input.

Our analysis above is based on the assumption that the arrivals to the virtual M/M/1/N queue of each output can be approximated by a Poisson process. To test the goodness of this assumption, we conducted goodness of fit tests

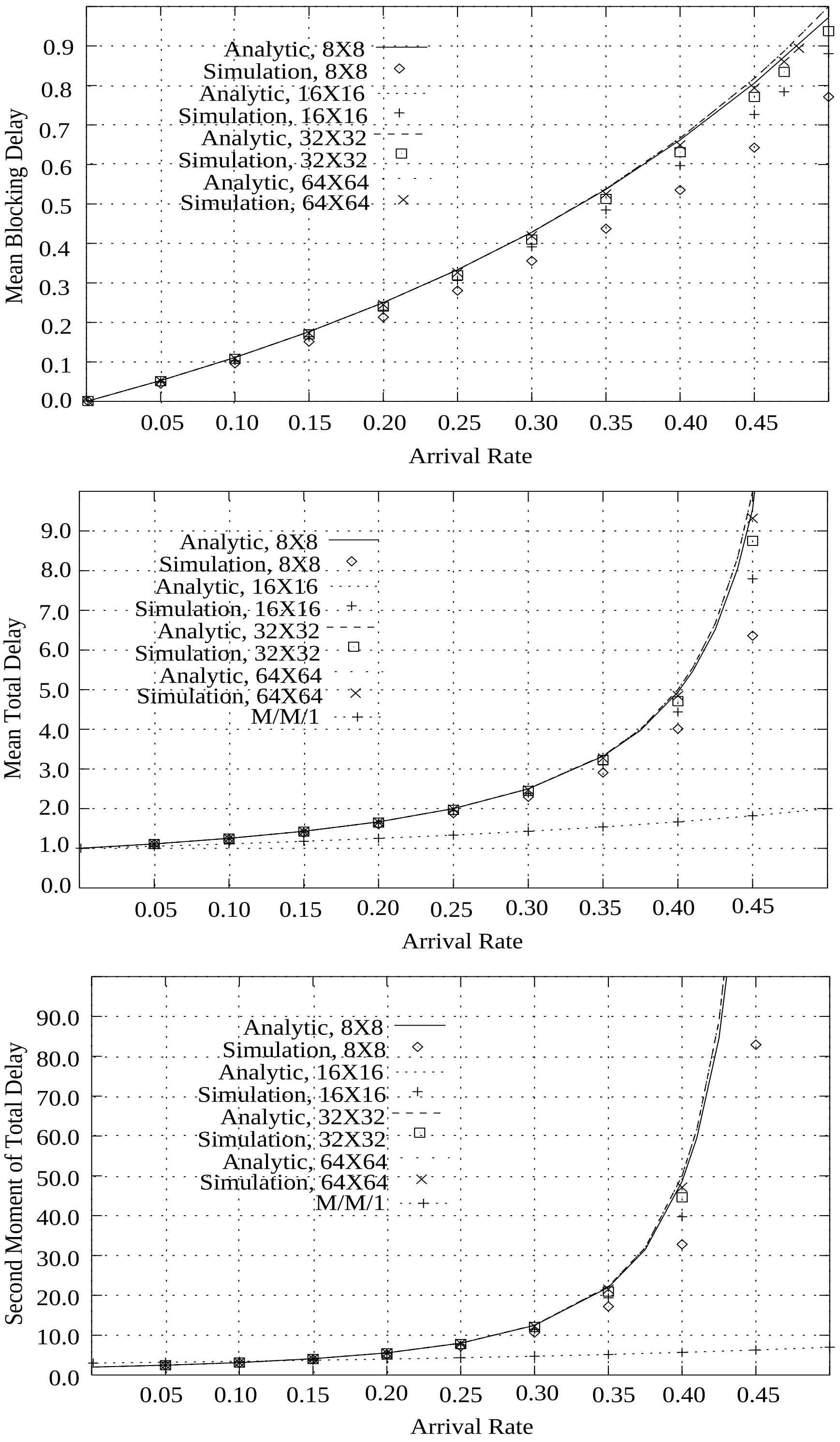


Fig. 3. First moment of blocking delay and the first and second moments of the waiting time *vs* throughput for Poisson arrivals and infinite buffers. From analytical and simulation models for $N = 8, 16, 32$ and 64.

for the packet interarrival times to the virtual queue having an exponential distribution. Three different tests were carried out for each scenario and we used the Kolmogorov, Cramér-von Mises and the Kuiper statistics [34] to test the hypothesis that the arrivals are from an exponential distribution. These tests are based on empirical distribution function (EDF) statistics and they are superior to Chi-square tests for continuous distributions. We used our simulator to get the timestamps at which packets destined for a particular output port moved to the HOL of its input port queue and thus "arrive" at the virtual M/M/1/N queue at the output port. Our test results show that even for switch sizes as small as 8×8, the interarrival times fit well with an exponential distribution for significance levels of 15%, 5% and 1%. Also, as the switch size increases, the significance levels of the tests statistics decrease, indicating that the interarrivals are an even closer match to an exponential distribution. To test for the Poisson nature, in addition to the tests for exponentiality, we need to test for the independence of the interarrivals. We carried out tests on the magnitude of the autocorrelation function at different lags, as per the procedure outlined in [30]. The magnitude of the autocorrelation function was smaller than the test statistic at all lags, validating the independence assumption. The detailed statistics are available in the technical report [20].

2.1 Finite Buffer Analysis

We now consider the case of finite buffers at each of the input ports. The analysis is similar to that for the infinite buffer case in that we first obtain the distribution of the time spent at the head of the queue and use this distribution in characterizing the input queue as an M/G/1/K queue with K buffers.

The "service time" distribution for the input M/G/1/K queue is obtained exactly like in the finite buffer case except that the throughput from output j will be $\Lambda_j(1-PB_j)$, where PB_j is the blocking probability of packets that were destined for output j. Now, as before, we approximate the arrival process to the virtual queue of output j by a Poisson process of throughput $\Lambda_j(1-PB_j)$ and model this queue as a M/M/1/M queue. Therefore, effective arrival rate Λ'_j, will be obtained by solving for Λ'_j in the equation

$$\Lambda_j(1-PB_j) = \Lambda'_j \left[\frac{1-{\eta'}_j^M}{1-{\eta'}_j^{M+1}} \right] \qquad (10)$$

where $\eta'_j = \Lambda'_j/\mu_j$. The Λ'_j obtained as above is used in Equations 3, 4 and 5 to calculate $\theta_j(k)$, $\pi_j(k)$ and $X_i(s)$ respectively and the expressions for the moments of the blocking delay and the "service time" are obtained from by

Equations 6 and 7 respectively.

The M/G/1/K_i queue with Poisson arrivals of rate λ_i and service time distribution given by Eqn. 5 for input port i is analyzed using the diffusion approximation method of [13]. From [13], the probability that there are n packets in input queue i, $\zeta_i(n)$, is given by

$$\zeta_i(n) = \begin{cases} c_i\hat{\zeta}_i(n), & 0 \leq n < K_i \\ 1 - \frac{1-c_i(1-\rho_i)}{\rho_i}, & n = K_i \end{cases} \tag{11}$$

where $\rho_i = \lambda_i\overline{X_i}$, $\hat{\zeta}_i(n)$ is the probability that there are n packets in an M/G/1/∞ with arrival rate λ_i and service time distribution corresponding to Eqn. 5, and

$$c_i = \left\{1 - \rho_i\left[1 - \sum_{n=0}^{K_i-1}\hat{\zeta}_i(n)\right]\right\}^{-1} \tag{12}$$

Using the diffusion approximation, the probabilities $\hat{\zeta}_i(n)$ can be written as

$$\hat{\zeta}_i(n) = \begin{cases} 1 - \rho_i, & n = 0 \\ \rho_i(1-\hat{\rho}_i)(\hat{\rho}_i)^{n-1} & n \geq 1 \end{cases} \tag{13}$$

where

$$\hat{\rho}_i = \exp\left(\frac{2\overline{X_i}^2(\lambda_i\overline{X_i} - 1)}{\lambda_i\overline{X_i}^3 + \overline{X_i^2} - \overline{X_i}^2}\right) \tag{14}$$

The mean delay through the switch for packets arriving into input i, D_i, is obtained by obtaining the mean queue length in input i and then applying Little's theorem.

$$D_i = \frac{1}{\lambda_i(1-\zeta_i(K_i))}\sum_{n=0}^{K_i} n\zeta_i n \tag{15}$$

Since the arrivals are Poisson, the blocking probability at input port i will be $\zeta_i(K_i)$.

Note that the effective arrival rate for an virtual queue of an output was derived using the blocking probability for packets destined for output port i, it is easy to see that

$$PB_j = \sum_{i=1}^{M} \zeta_i(K_i)\frac{p_{ij}\lambda_i}{\Lambda_j} \tag{16}$$

The analytical model is solved by iterating on Eqns 1-16. For a given arrival rate, we start with an arbitrary assumption for the blocking probability and calculate the effective arrival rate using Eqn /refeqn-Lambda-j-prime-finite. Eqns 3-7 and 11-16 are then used to obtain the new blocking probability. This procedure is repeated till the blocking probabilities from successive iterations have a difference of less than some prespecified error margin. For our calculations, we stopped iterating when the values from two successive iterations was less that 0.0001.

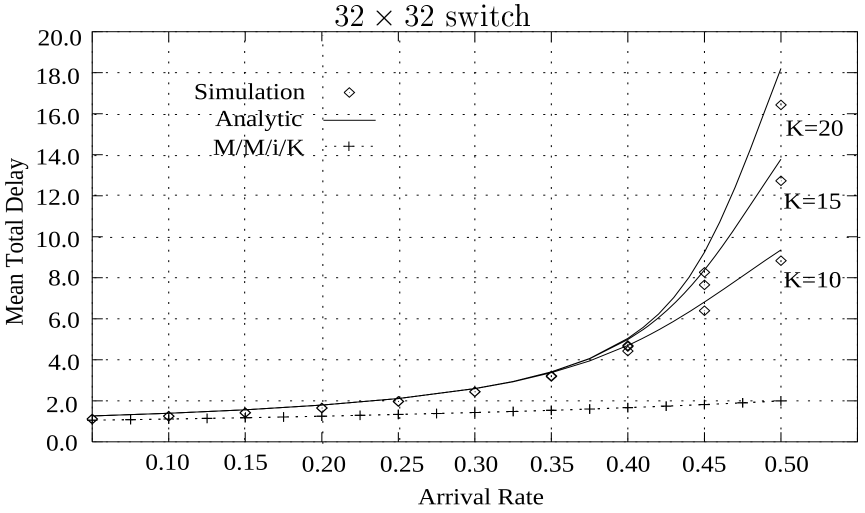
For the numerical results we consider 32 × 32 and 64 × 64 switches with identical arrival rates at all the inputs and uniform routing probabilities. We obtain the mean delay and the blocking probabilities for different arrival rates and buffer sizes. To study the goodness of our approximation, we compare the analytical results with a simulation model that is identical to the one described in Section 2 except that now we consider fnite buffers rather than infinite buffers.

In Fig. 4 we show the mean queuing delay. Note the good agreement between the simulation and the analytical results that validates our approximation. In these figures we also show the mean delay in an output queued switch, we can model as an M/M/1/K queue. Note the considerably high "delay-penalty" for the input queued switch.

In Tables 1 and 2 we show the loss probabilities for different arrival rates and buffer sizes. Once again, we remark on the goodness of our approximation as seen by the close match between the analytical and simulation results. In these tables we also show the blocking probabilities for a pure output queued switch, in which case each output queue can be modeled as an M/M/1/K queue. We observe that there is considerable penalty in input input queueing and the blocking probability can increase by many orders of magnitude.

	K=10			K=15			K=20		
λ	Sim.	Ana.	o/p queue	Sim.	Ana.	o/p queue	Sim.	Ana.	o/p queue
0.40	0.0048	0.0059	6.30E-5	0.0006	0.0008	6.44E-7	0.0001	0.0001	6.60E-9
0.50	0.0571	0.0650	0.0005	0.0363	0.0443	1.53E-5	0.0256	0.0335	4.77E-7
0.60	0.1702	0.1815	0.0024	0.1595	0.1718	0.0002	0.1560	0.1687	1.46E-5

Table 1
Loss probabilities in a 32×32 switch with Poisson arrivals



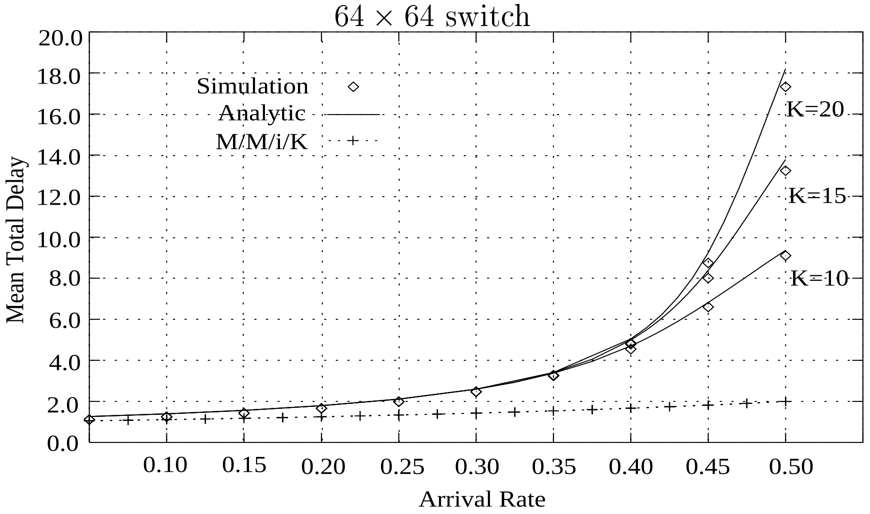


Fig. 4. First moment of total delay *vs* throughput for Poisson arrivals and finite buffers. From analytical and simulation models for $N = 32$ and 64.

	K=10			K=15			K=20		
λ	Sim.	Ana.	o/p queue	Sim.	Ana.	o/p queue	Sim.	Ana.	o/p queue
0.40	0.0052	0.0059	6.30E-5	0.0007	0.0008	6.44E-7	0.0001	0.0001	6.60E-9
0.50	0.0610	0.0650	0.0005	0.0401	0.0443	1.53E-5	0.0295	0.0336	4.77E-7
0.60	0.1759	0.1815	0.0024	0.1657	0.1718	0.0002	0.1622	0.1687	1.46E-5

Table 2
Loss probabilities in a 64×64 switch with Poisson arrivals

3 $M \times N$ Switch with Self Similar Input and Exponential Packet Lengths

Having modeled switch behavior under the somewhat idealized model of Poisson inputs we will now examine the behavior under a more realistic model of self similar inputs. Before presenting the delay analyses for self similar arrival processes we give a brief overview of the various equivalent definitions of self similarity and the packet arrival models that can be used with each of these. Finally, we will select that definition and packet arrival model that has a well developed queuing theory.

Packet arrival instants are modeled as point processes. Divide the time axis into non overlapping intervals of unit length and let $X = \{X_t : t = 0, 1, 2, \cdots\}$ be the number of points (packet arrivals) in the t^{th} interval. Measurements and analysis of such packet arrival processes in real networks has indicated that X is a self similar process. This means that although the analysis of packet switches for the Poisson packet arrival model gives us a "first-order-feel" for their performance, to understand their performance in real networks, it is necessary to study their performance for self similar packet arrivals.

Mathematically, self similarity in the process X can be expressed in many ways. Let X be covariance stationary with mean λ, variance σ^2 and autocorrelation function $r(k), k \geq 0$. For each $m = 1, 2, \cdots$, let $X^{(m)} = (X_k^{(m)} : k = 1, 2, \cdots)$ be the new covariance stationary time series (with corresponding autocorrelation function $r^{(m)}$) obtained by averaging the original series X over non-overlapping blocks of size m, i.e., for each $m = 1, 2, \cdots$, $X^{(m)} = (X_{km-m+1} + \cdots + X_{km})/m$, $k \geq 1$. Then self-similarity of X means any of the following

X has a slowly decaying variance: The variance of the sample mean decreases more slowly than the reciprocal of the sample size. $\mathsf{var}(X^{(m)}) \sim am^{-\beta}$ as $m \to \infty$ and $0 < \beta < 1$ (a is a finite positive constant)

X is long range dependent (LRD): The autocorrelations decay hyperbolically rather than exponentially fast, implying a non-summable autocorrelation function. $\sum_k r(k) = \infty$ and

X is $1/f$-noise: The spectral density $f(\cdot)$ obeys a power-law near the origin, i.e., $f(\lambda) \sim b\lambda^{-\gamma}$, as $\lambda \to 0$ with $0 < \gamma < 1$ and $\gamma = 1 - \beta$ (b is a finite positive constant)

Each of the above descriptions of a self similar process can lead to a class of models for the packet arrival process. From the point of understanding queuing behavior of systems, we consider those that are derived to match the LRD statistics of the packet arrival process. In [18] Leland *et al* show that Gaussian noise or nonlinear transformations on Gaussian noise such as fractional ARIMA can be used to characterize a LRD X. In [30], Paxson and Floyd show that superposition of on/off sources that have a fixed rate in the on period and have a heavy-tailed distribution for the on and off period lengths can be used to model LRD X. Erramilli, Singh and Pruthi use deterministic nonlinear chaotic maps to define a LRD X [9]. Andersen and Nielsen propose a Markovian approach in which an LRD X is obtained by superposing a number of two state Markov Modulated Poisson Processes (MMPPs) [1] with the resultant arrival process being an MMPP. The advantage of this last method is that in addition to allowing the modeling of burstiness over a number of time scales with the desired covariance structure, since the packet arrivals are MMPP, a well developed queuing theory is available for analysis. Additionally, our choice of the MMPP arrival process to model LRD packet arrival processes is motivated by the arguments in [14] where it is shown that in practical networking scenarios where we have finite buffer sizes, we need to account for the correlation over only for a finite number of time scales corresponding to the buffer size and a Markovian approach can be used to obtain accurate performance predictions since a power law decay can be approximated arbitrarily closely by enough exponential decay functions. Further, in [33] it is shown that the MMPP model converges to fractional Brownian motion under limiting conditions. The convergence is in the sense of weak convergence implying the same queueing behavior as with fractional Brownian motion. By increasing the number of MMPPs in the superposition, the long-range dependent correlation structure can be captured over an arbitrarily large number of time scales. Therefore, we will use this in the analysis of the variable length packet switches with input queuing.

We first summarize the technique outlined in [1] to fit an MMPP process to an LRD arrival process. Let the packet arrival process to input port i be a second order self similar process with mean λ_i, correlation at lag 1 ρ_i, Hurst parameter H_i, and the number of time scales over which the burstiness is to be modeled, n_i. This will be modeled as the superposition of a number of two state Interrupted Poisson Processes (IPPs), typically four, and a Poisson process. The covariance function of this superposed process is fitted to that of the self-similar process that we are modeling over several time scales. For input port i we will superpose d_i IPPs and the jth two-state IPP is parameterized by its generator matrix Q_i^j and rate matrix R_i^j as follows

$$Q_i^j = \begin{bmatrix} -c_i^{1j} & c_i^{1j} \\ c_i^{2j} & -c_i^{2}j \end{bmatrix} \qquad R_i^j = \begin{bmatrix} r_i^j & 0 \\ 0 & 0 \end{bmatrix} \tag{17}$$

The superposed process will be

$$\mathbf{Q}_i = \bigoplus_{j=1}^{d_i} Q_i^j \qquad \mathbf{R}_i = \bigoplus_{j=1}^{d_i} R_i^j \tag{18}$$

where $\bigoplus$ denotes the Kronecker sum. Note that the individual Poisson process in the fitting procedure may also be represented as a special case of a MMPP and added in the Kronecker sum of Eqn 18 to obtain the complete MAP model of the arrival process. The steady state probability vector of the Markov chain,

$\mathbf{\Phi}_i$, can be obtained by simultaneously solving the following equations,

$$\mathbf{\Phi}_i \mathbf{Q}_i = 0 \qquad \mathbf{\Phi}_i \mathbf{e}_i = 1 \tag{19}$$

where $\mathbf{e}_i = [1, 1, \cdots, 1]^T$ is a unit column vector of length 2^{d_i}. Let $\mathbf{r}_i = [r_i^1, r_i^2, \cdots, r_i^{d_i}]$. Then the average arrival rate to input i, $\lambda_i = \mathbf{\Phi}_i \mathbf{r}_i^T$. The procedure to fit c_i^{1j}, c_i^{2j} and r_i^j to λ_i, ρ_i, H_i and n_i is described in [1].

As in the previous section we assume that each packet at input i chooses output j independent of other packets with probability p_{ij} and the rate at which a packet is evacuated from an input queue by output port j is μ_j which is the line rate at output port j. Packet lengths are exponentially distributed with unit mean. There are infinite buffers at the input and none at the output. The output ports evacuate packets from the HOL of the input queues according to “first blocked first served” discipline. The “service time” of the input queue, time spent at the HOL by packet, is obtained exactly as before by making the approximation that the virtual queue to each output is an M/M/1/M queue. The sojourn time in this M/M/1/M queue is thus the service time for the input queue which we can now model as an MMPP/G/1 queue. Since the service time for the input queue is like before, the maximum throughput per port will be 0.5 and is derived exactly as before. Thus the moments of the service times are obtained exactly like in the previous section using Eqns 1-7. The first and second moments of the packet delays in the input queue can now be obtained using well known techniques for MMPP/G/1 queues [11]. The procedure is summarized in the appendix. We also note that statistical goodness of fit tests were carried out on the interarrival times at the virtual M/M/1/N to test for their exponential nature. As in Section 2, we used the Kolmogorov, Cramér-von Mises and the Kuiper statistics and the results verified that the interarrivals times are indeed from an exponential distribution. We also carried out tests for independence of the interarrivals by checking the magnitude of the autocorrelation function and validated the independence assumption. Once again, the test results are available in [20].

Numerical results are obtained as follows. We use the Bellcore traces [18] and derive their statistical properties in terms of the Hurst parameter, the correlation at lag 1 and the time scales over which the burstiness occurs. These parameters and the arrival rate λ are used to fit the parameters c_i^{1j}, c_i^{2j} and r_i for $j = 1, \cdots, 4$ of the MMPP model described in [1]. The analytical results are obtained for the MMPP/G/1 queue as described earlier. To validate the analytical results we also develop a simulation model in which the arrivals are MMPP with parameters derived above. The arrival process generator is validated by simulating a single server queue and comparing with the results given in [10]. The magnitudes of our delays and the knee region of the delay-throughput graph match that given in Figure 2 of [10]. In the simulation model a separate and independent MMPP arrival process generator is used

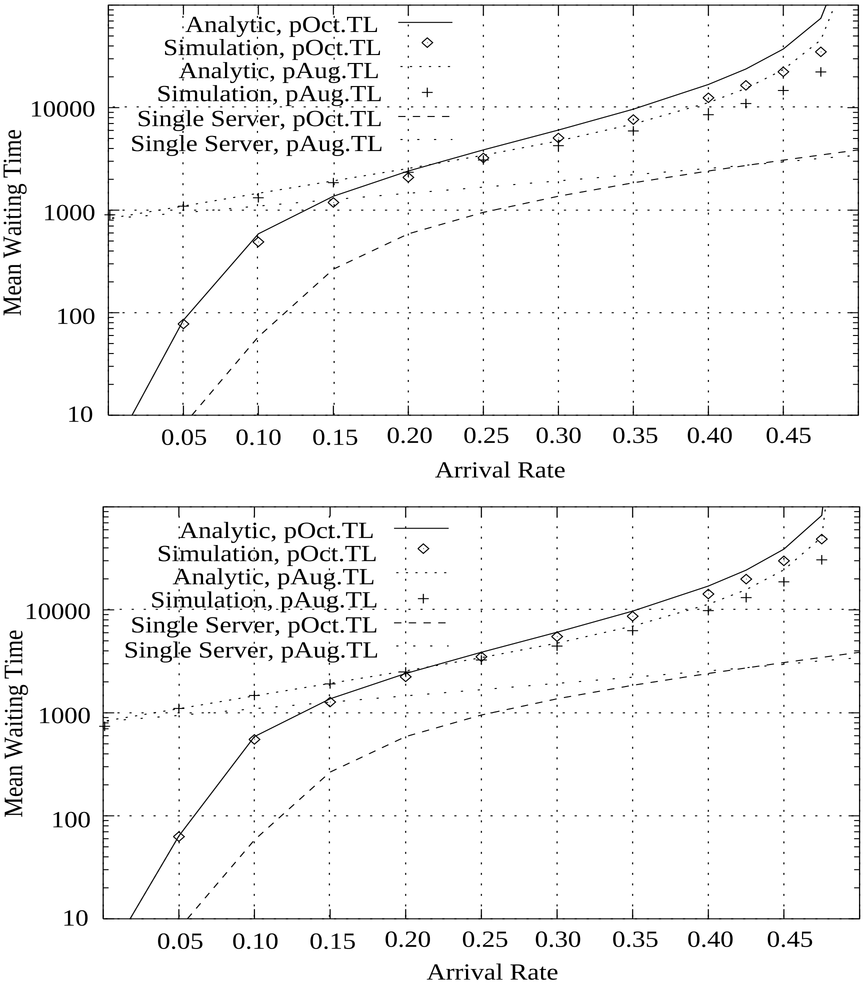


Fig. 5. First moment of total delay *vs* throughput for self similar arrivals and infinite buffers. Results are from analytical and simulation models for Bellcore traces `pAug.TL` and `pOct.TL`. Top graph shows results for an 8 × 8 switch and bottom graph for a 16 × 16 switch.

for each of the input ports with the traces generated by each of the sources having identical statistical properties. Thus, statistically identical self-similar traces but with different sample paths are used as the input processes to the simulation model. In this paper we primarily use the Bellcore traces `pAug.TL` ($H = 0.82$ and $\rho = 0.582$) and `pOct.TL` ($H = 0.92$ and $\rho = 0.356$). We model burstiness over 4 time scales.

We mention here that we considered feeding the traces to obtain the simulation results. Since the number of inputs was large, the size of the traces was insufficient. The same trace cannot be fed to all the inputs because in that case the arrivals at each input will have a correlation of one, an obviously wrong choice for an arrival process. Also, we did not use shuffled versions of a single trace because shuffling of the time series of the traces would lead to a loss of the correlation structure and consequently the long range dependence.

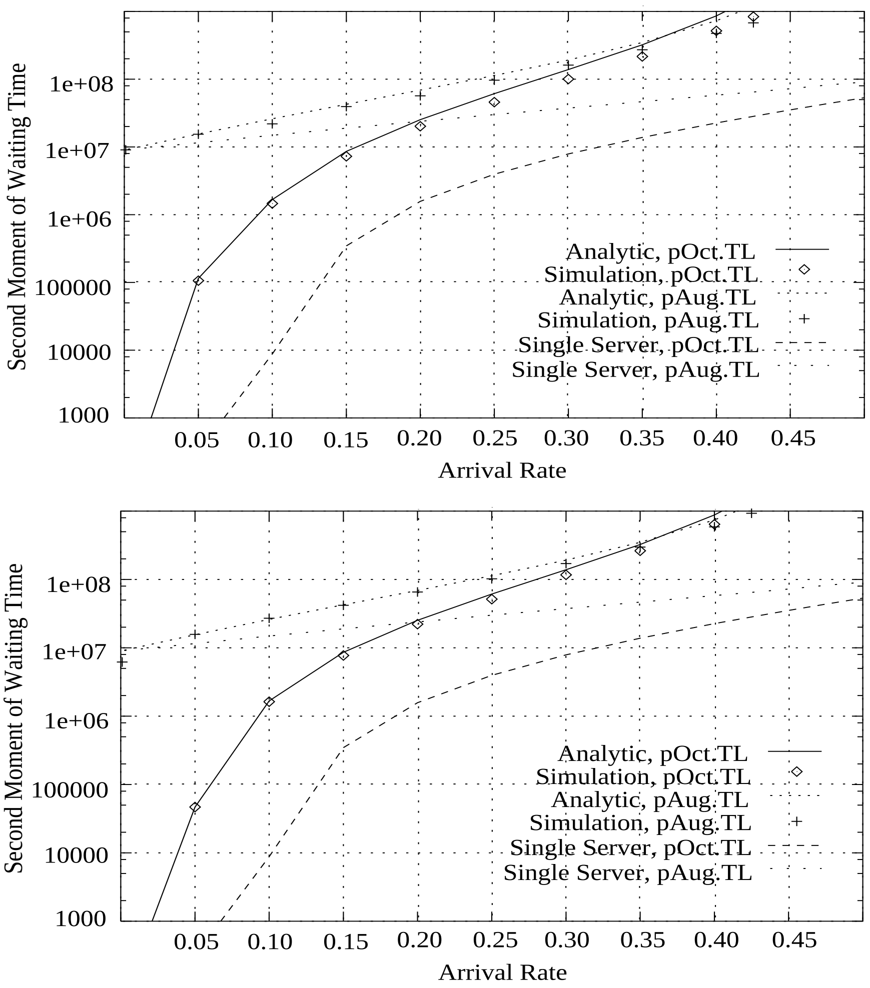


Fig. 6. Second moment of total delay *vs* throughput for self similar arrivals and infinite buffers. From analytical and simulation models for Bellcore traces `pAug.TL` and `pOct.TL`. Top graph shows results for an 8 × 8 switch and bottom graph for a 16 × 16 switch.

In Figs. 5–8 we show the first and second moments of total and blocking delays in the switch. It can be seen that the simulation and analytical results are in extremely good agreement except at loads close to the capacity of the switch. We see a marked difference in the shape of the delay characteristics for the `pOct.TL` trace at low loads which can be attributed to its comparatively low correlation value at lag one. At low loads, the low correlation suggests a lower probability of successive intervals having packet arrivals, which in turn leads to low delays. Further investigation of the effect of the correlation structure is done in Section 6. As discussed earlier, the throughput delay curves in Figure 5 show that the switch saturates at a load of 0.5. Also, note that the first and second moments of the blocking delay shown in Figures 7 and 8 are identical for both the traces for a given switch size. This is because the virtual queue at each output port is modeled as an M/M/1/M queue whose delay characteristics depend only on the average arrival rate of the input processes

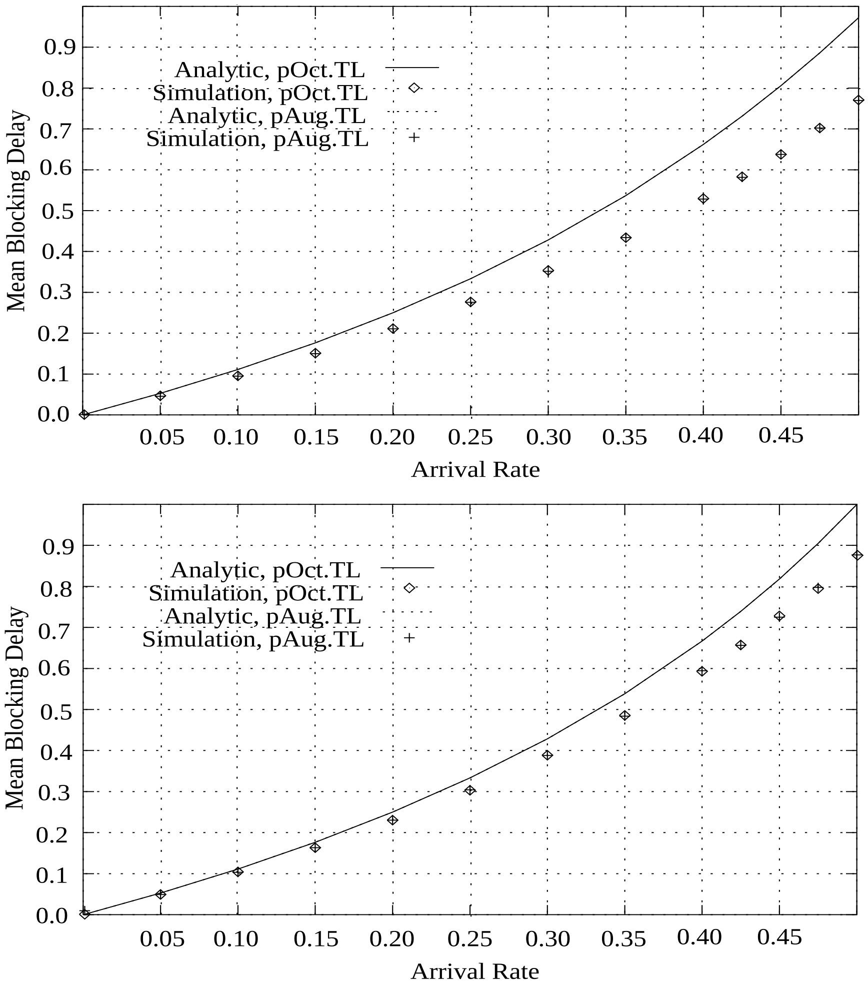


Fig. 7. First moment of blocking delay *vs* throughput. for self similar arrivals and infinite buffers. Results from analytical and simulation models for Bellcore traces `pAug.TL` and `pOct.TL`. Top graph shows results for an 8×8 switch and bottom graph for a 16×16 switch.

and not on any of their other statistical properties.

From Figure 5 we see that the mean delay increases exponentially as the arrival rate increases. The delay performance can be divided into three regions - low $(0.0 - 0.10)$, medium $(0.10 - 0.40)$ and high $(0.40 - 0.50)$ loads. Note that in the medium load region the mean delay is of the order of the order of 10^3. In all these regions the mean delay increases exponentially with increasing arrival rate. For comparison, we have shown the delays that would have been experienced in a single server queue without HOL blocking. This would be the delay experienced in an output queued switch in which the arrival rate to an output port would be described by the corresponding MMPP process. This shows that for a given arrival rate mean delay in the input queued switch could be at least double and nearly 10 times higher even at medium load.

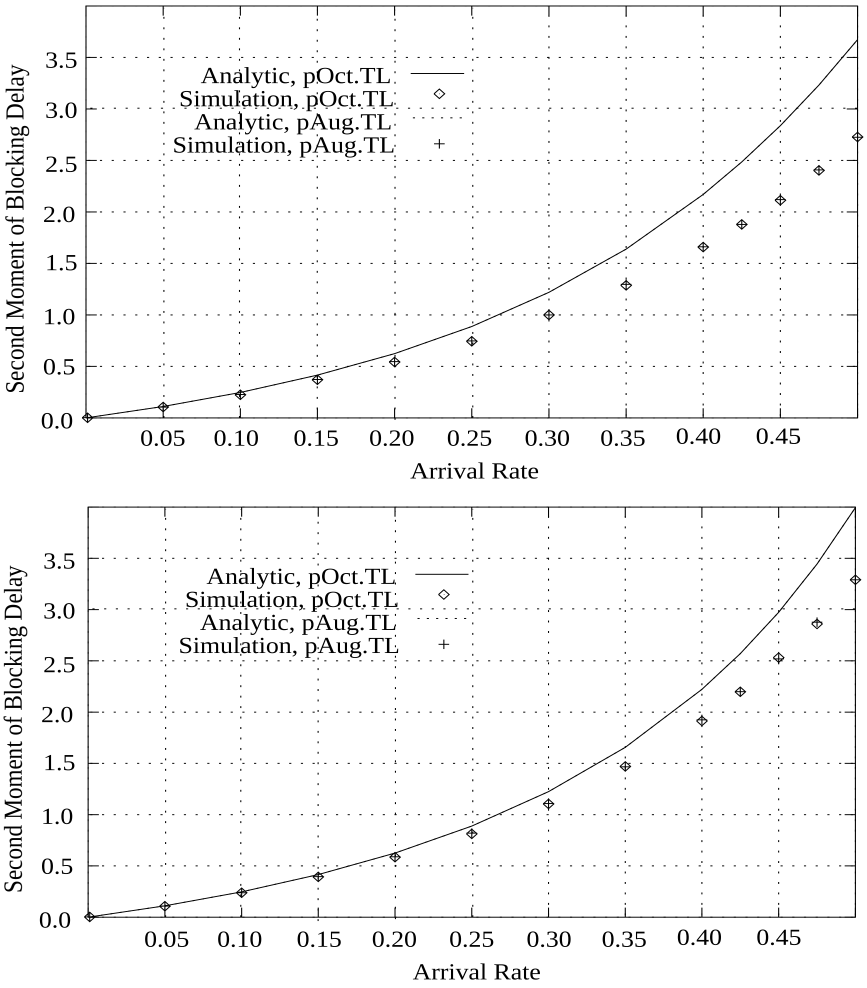


Fig. 8. Second moment of blocking delay *vs* throughput for self similar arrivals and infinite buffers. From analytical and simulation models for Bellcore traces `pAug.TL` and `pOct.TL`. Top graph shows results for an 8×8 switch and the bottom graph for a 16×16 switch.

The moments of the blocking delay for the case of Poisson arrivals and that of the MMPP arrivals is identical in the analytical models. Comparisons with the simulation model suggests that the analytical models are a good approximation. Hence we note that the effect of increase in the second moment in the case of self similar arrivals is significantly larger.

We have performed extensive analysis and simulations to understand the switch behavior under self similar arrivals and we have observed that when the burstiness extends over 3 time scales, the delays are of the order of 10^2.

From the above results we note that the analytical results match the simulations reasonably well. Therefore, in the following we do not present any simulation results.

We only report the loss analysis for the MMPP arrival process. Having obtained the service time distribution and described the arrival process we use the MMPP/G/1/K analysis of [3] and the efficient evaluation techniques for evaluating the loss probabilities from [4] to solve our model. The following notation will be used for each input i and to simplify the notation we will omit the subscript corresponding to the input port.

$\mathbf{U}$: $m \times m$ matrix given by $(\mathbf{R} - \mathbf{Q})^{-1}\mathbf{R}$

$\mathbf{P}(n,t)$: $m \times m$ matrix whose (p,q)th element denotes the conditional probability of reaching phase q of the MMPP and having n arrivals during a time interval of length t, given that we start with phase p at time $t = 0$.

$\mathbf{A}_n$: $m \times m$ matrix given by $\int_0^\infty \mathbf{P}(n,t)dX(t)$, $n \geq 0$ where $X(t)$ is the distribution of the service time as obtained previously.

$\mathbf{A}$: $m \times m$ matrix given by $\sum_{n=0}^{\infty} \mathbf{A}_n = \int_0^\infty e^{\mathbf{R}t}dX(t)$

$\mathbf{\Pi}(i)$: m dimensional vector whose p^{th} element is the limiting probability at the embedded epochs of having i packets in the queue and being in phase p of the MMPP, $i = 0, 1, \cdots, K-1$.

$\mathbf{I}$: $m \times m$ identity matrix

The steady-state probability distribution of the queue length of the embedded Markov chain at the departure instants can be calculated using the following approach. The matrices $\mathbf{A}_n$ are first calculated using the technique described in [11]. From $\mathbf{P}$, we then find the matrix sequence $\{\mathbf{C}_i\}$, independent of the buffer size K, such that $\mathbf{\Pi}(i) = \mathbf{\Pi}(0)\mathbf{C}_i$ for $i = 0, 1, \cdots, K-1$. The matrices $\mathbf{C}_i$ are calculated using the following equation

$$\mathbf{C}_{i+1} = \left[\mathbf{C}_i - \mathbf{U}\mathbf{A}_i - \sum_{v=1}^{i} \mathbf{C}_v \mathbf{A}_{i-v+1}\right] \mathbf{A}_0^{-1} \quad i = 1, \cdots, K-2 \tag{20}$$

beginning with $\mathbf{C}_0 = \mathbf{I}$. The vector $\mathbf{\Pi}(0)$ is then determined using

$$\mathbf{\Pi}(0)\left[\sum_{v=0}^{K-1} \mathbf{C}_v + (\mathbf{I} - \mathbf{U})\mathbf{A}(\mathbf{I} - \mathbf{A} + \mathbf{e}\mathbf{\Phi})^{-1}\right] = \mathbf{\Phi} \tag{21}$$

The loss probability can then be found using the following expression

$$\mathrm{P}_{\text{loss}} = 1 - (\overline{X}\mathbf{\Phi}\mathbf{R}\mathbf{e})^{-1}\left[1 + \mathbf{\Pi}(0)(\mathbf{R} - \mathbf{Q})^{-1}\overline{X}^{-1}\mathbf{e}\right]^{-1} \tag{22}$$

Numerical results are obtained as follows. We use the Bellcore traces [18] and derive their statistical properties in terms of the Hurst parameter, the

λ	K=250			K=500			K=1000		
	Sim.	Ana.	o/p queue	Sim.	Ana.	o/p queue	Sim.	Ana.	o/p queue
0.10	0.0662	0.0669	0.0506	0.0567	0.0649	0.0488	0.0541	0.0610	0.0454
0.15	0.0764	0.0812	0.0549	0.0672	0.0790	0.0531	0.0629	0.0749	0.0498
0.20	0.0977	0.1014	0.0591	0.0760	0.0990	0.0573	0.0746	0.0944	0.0540

Table 3
Loss probabilities in a 8 × 8 switch for self similar traffic corresponding to the Bellcore trace `pAug.TL`

λ	K=250			K=500			K=1000		
	Sim.	Ana.	o/p queue	Sim.	Ana.	o/p queue	Sim.	Ana.	o/p queue
0.10	0.0260	0.0363	0.0006	0.0216	0.0297	0.0003	0.0143	0.0198	6.96E-5
0.15	0.0598	0.0769	0.0134	0.0553	0.0690	0.0090	0.0414	0.0556	0.0041
0.20	0.0928	0.1188	0.0321	0.0857	0.1103	0.0266	0.0743	0.0951	0.0183

Table 4
Loss probabilities in a 8×8 switch for self similar traffic corresponding to the Bellcore trace `pOct.TL`

correlation at lag 1 and the time scales over which the burstiness occurs. These parameters and the arrival rate λ are used to fit the parameters c_i^{1j}, c_i^{2j} and r_i for $j = 1, \cdots, 4$ of the MMPP model described in [1]. The analytical results are obtained for the MMPP/G/1 queue as described earlier. To validate the analytical results we also develop a simulation model in which the arrivals are MMPP with parameters derived above. The arrival process generator is validated by simulating a single server queue and comparing with the results given in [10]. The magnitudes of our delays and the knee region of the delay-throughput graph match that given in Figure 2 of [10]. In the simulation model a separate and independent MMPP arrival process generator is used for each of the input ports with the traces generated by each of the sources having identical statistical properties. Thus, statistically identical self-similar traces but with different sample paths are used as the input processes to the simulation model. In this paper we primarily use the Bellcore traces `pAug.TL` ($H = 0.82$ and $\rho = 0.582$) and `pOct.TL` ($H = 0.92$ and $\rho = 0.356$). We model burstiness over 4 time scales.

Tables 3 and 4 present the overflow probabilities for a 8 × 8 switch fed with the Bellcore traces `pAug.TL` and `pOct.TL`. Note that loss probabilities are high even for low loads and that the drop in the loss rates does not scale with increase buffer sizes. Such behavior of queues fed with long-range dependent traffic have also been reported in [10]. Thus large number of buffers are required with long-range dependent traffic to support even moderate loss probabilities. The tables also show the loss probabilities for the MMPP/M/1/K

queue for different values of K. Bernoulli splitting of the packet arrivals into N streams corresponding to splitting the packet arrivals to the N outputs and the combining of M such streams at the output queue will make the characteristics of the arrival process to the output queue different from that to the input queue. However, we just use it as an approximation to get an indication of the "input queueing penalty" for the more realistic model of self-similar packet arrivals.

4 Increasing Throughput in Input Queued Switches

Recall that we suggested that a FIFO-CIOQ with speedup and/or parallelism would be architecturally simpler than a VOQ-CIOQ switch, especially for variable length packets. In this section we provide the throughput delay analysis of such a switch. It is easy to see that in our analysis speedup is modeled by using a higher μ for the evacuation rate, or equivalently a lower λ for the arrival rate. In the following we analyze a switch with parallelism of m, $m > 1$.

4.1 Delay Analysis for Parallelism of m, $m > 1$

The delay analysis is along the same lines as the previous analyses. First, let us consider the infinite buffer case. It is easy to see that in this case if there are more than m HOL packets at the inputs destined for a particular output port, m of them are served simultaneously while the others are blocked. Here too we assume the input process to the queue to be Poisson which is an approximation when M is finite. Thus the virtual queue of each output port will be modeled as an M/M/m/M queue and the effective arrival rate to output port j corresponding to a throughput of Λ_j is obtained by solving for Λ'_j in

$$\Lambda_j = \Lambda'_j \left[1 - \frac{\left[\frac{(\eta'_j)^M m^m}{m!} \right]}{\left[\sum_{k=0}^{m-1} \frac{(m\eta'_j)^k}{k!} + \sum_{k=m}^{M} \frac{(\eta'_j)^k m^m}{m!} \right]} \right] \tag{23}$$

where $\eta'_j = \frac{\Lambda'_j}{m\mu_j}$. $\theta_j(k)$, $\pi_j(k)$ and $X_i(s)$ are obtained like before by considering an M/M/m/M queue rather than an M/M/1/M queue at the outputs. Similarly the blocking and total service time moments are also obtained like before and are given by,

$$\overline{B_i} = \sum_{j=1}^{N} p_{ij} \sum_{k=m}^{M-1} \pi_j(k) \frac{k - m + 1}{\mu_j}$$

$$\overline{B_i^2} = \sum_{j=1}^{N} p_{ij} \sum_{k=m}^{M-1} \pi_j(k) \frac{(k-m+1)(k-m+2)}{\mu_j^2}$$

$$\overline{B_i^3} = \sum_{j=1}^{N} p_{ij} \sum_{k=m}^{M-1} \pi_j(k) \frac{(k-m+1)(k-m+2)(k-m+3)}{\mu_j^3} \quad (24)$$

$$\overline{X_i} = \overline{B_i} + \sum_{j=1}^{N} \frac{p_{ij}}{\mu_j}$$

$$\overline{X_i^2} = \overline{B_i^2} + 2\overline{B_i} \sum_{j=1}^{N} \frac{p_{ij}}{\mu_j} + 2 \sum_{j=1}^{N} \frac{p_{ij}}{\mu_j^2}$$

$$\overline{X_i^3} = \overline{B_i^3} + 3\overline{B_i^2} \sum_{j=1}^{N} \frac{p_{ij}}{\mu_j} + 6\overline{B_i} \sum_{j=1}^{N} \frac{p_{ij}}{\mu_j^2} + 6 \sum_{j=1}^{N} \frac{p_{ij}}{\mu_j^3} \quad (25)$$

Note that the summations over the index k for the blocking delay is from $k = m$ to $k = M - 1$ because only when there are m or more packets waiting in the virtual queue will the packet at the HOL of an input queue have to wait. We can now use the expressions for the average delay and its second moment as given in Eqns 8 and A.1 to obtain the latency for the case of Poisson and self similar arrivals respectively.

We show numerical results only for the self similar arrival case. Figure 9 shows the analytical results for the delay throughput characteristics for $N \times N$ switches with $N = 8, 16, 32$ and 64 for parallelism of 2 and 4. We assume identical loads on all the inputs and uniform routing probabilities p_{ij}. We see that effect of the switch size on the delay characteristics becomes negligible as the switch size increases. Also, the medium load region can be extended till the arrival rate of 0.75 for a parallelism of 2 and up to 0.85 for a parallelism of 4. Further, the mean delay is considerably lower with parallelism than without. Also, the steep rise in the mean delay in the low load region does not manifest in the speeded up switch.

The maximum throughputs for a given parallelism is obtained by solving for λ in $\lambda\overline{X} = 1.0$, where $\overline{X}$ is obtained from Eqn 25. Table 5 shows the maximum achievable throughputs for switches of various sizes and for parallelism of 2, 3 and 4. Note that a switch with a parallelism of 4 can support loads in excess of 99%.

The analysis can be extended to the case of finite input buffers. We first consider the case of Poisson arrivals. The input queues become M/G/1/K queues with the service time moments obtained from Eqn 25. In Fig. 10 we show the throughput-mean delay characteristics for a CIOQ switch with Poisson arrivals, and parallelism of 2 and 4 for various arrival rates normalized to the

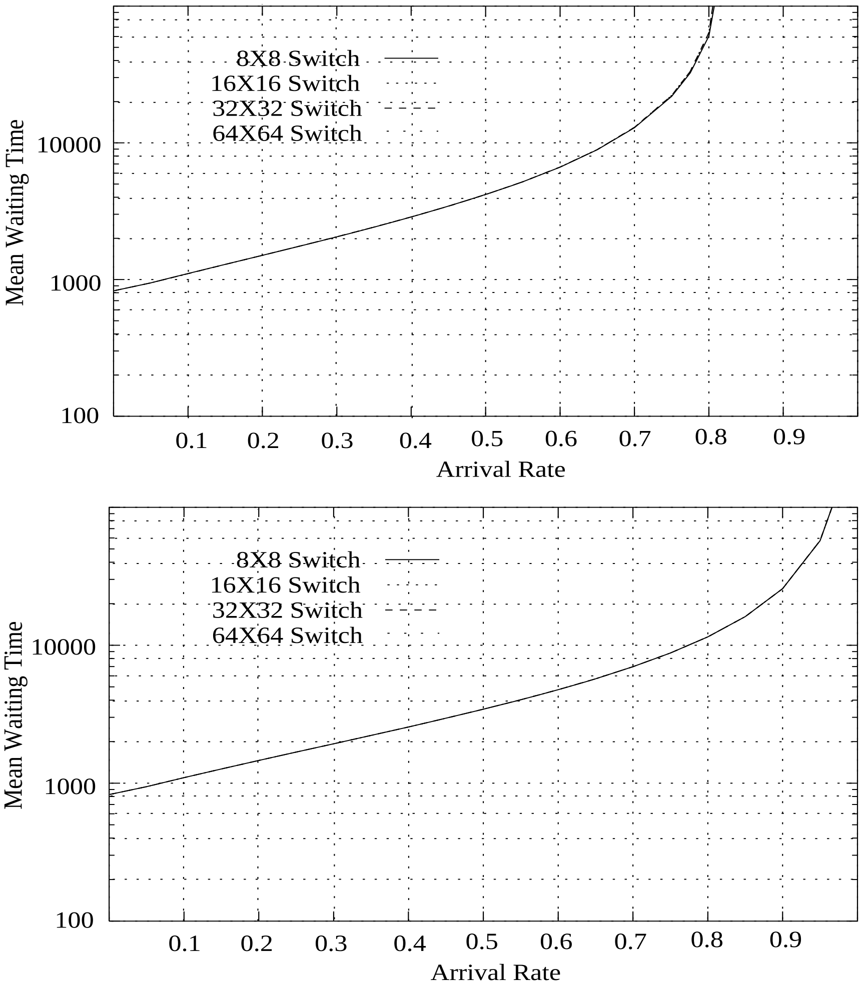


Fig. 9. First moment of total delay *vs* throughput for self similar arrivals and infinite input buffers. From analytical model for Bellcore trace `pAug.TL` for parallelism of 2 and 4. Results are shown for $N \times N$ switches with $N = 8, 16, 32$ and 64.

	Parallelism		
N	2	3	4
4	0.8670	0.9795	1.0000
8	0.8304	0.9616	0.9934
16	0.8284	0.9611	0.9934
32	0.8284	0.9611	0.9934
∞	0.8284	0.9611	0.9934

Table 5
Maximum throughput for various parallelism.

switch rate. A 70% load with a parallelism of 2 corresponds to an arrival rate of 0.35 and a parallelism of 4 corresponds to an arrival rate of 0.175. Note that the delay in the input queue is virtually negligible. In Tables 6 and 7 we

λ	K=10			K=15			K=20		
	Sim.	Ana.	o/p queue	Sim.	Ana.	o/p queue	Sim.	Ana.	o/p queue
0.60	0.0046	0.0048	0.0024	0.0005	0.0006	0.0002	0.0001	0.0001	1.46E-5
0.70	0.0196	0.0202	0.0086	0.0057	0.0060	0.0014	0.0017	0.0018	0.0002
0.80	0.0571	0.0590	0.0235	0.0333	0.0352	0.0072	0.0214	0.0233	0.0023

Table 6
Loss probabilities in a 64×64 switch with a parallelism of 2. Poisson arrivals.

show the blocking probabilities for different arrival rates and parallelism of 2 and 4 for a 32×32 switch. For the case of self similar arrivals we only report the loss probabilities in Tables 8 and 9 like in Section 3.1.

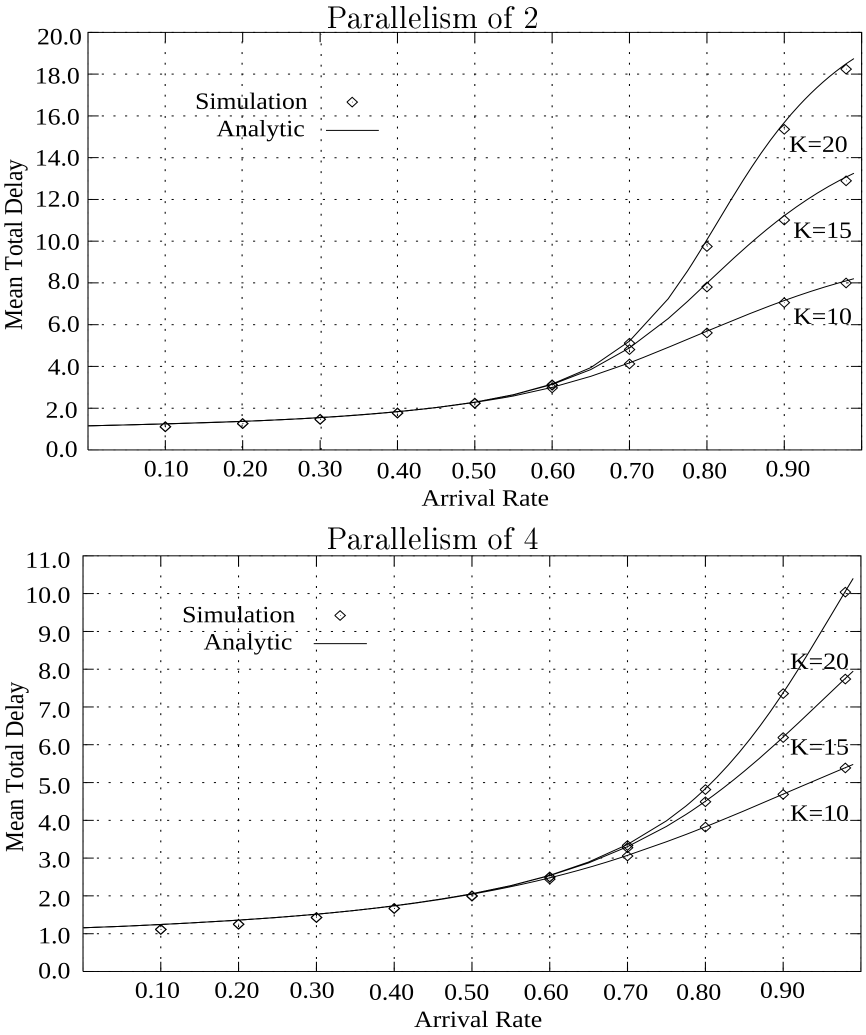


Fig. 10. First moment of total delay for a 32×32 switch with Poisson arrivals, finite buffers with link parallelism of 2 and 4. results from analysis and simulation

λ	K=10			K=15			K=20		
	Sim.	Ana.	o/p queue	Sim.	Ana.	o/p queue	Sim.	Ana.	o/p queue
0.60	0.0024	0.0026	0.0024	0.0002	0.0002	0.0002	1.40E-5	1.80E-5	1.46E-5
0.70	0.0087	0.0090	0.0086	0.0015	0.0015	0.0014	0.0003	0.0003	0.0002
0.80	0.0238	0.0240	0.0235	0.0074	0.0075	0.0072	0.0024	0.0024	0.0023

Table 7
Loss probabilities in a 64 × 64 switch with a parallelism of 4. Poisson arrivals.

λ	K=250			K=500			K=1000		
	Sim.	Ana.	o/p queue	Sim.	Ana.	o/p queue	Sim.	Ana.	o/p queue
0.10	0.0573	0.0505	0.0506	0.0516	0.0487	0.0488	0.0444	0.0453	0.0454
0.15	0.0643	0.0549	0.0549	0.0523	0.0531	0.0531	0.0513	0.0497	0.0498
0.20	0.0686	0.0596	0.0591	0.0595	0.0579	0.0573	0.0559	0.0545	0.0540

Table 8
Loss probabilities in a 8 × 8 switch with a parallelism of 2 and self similar arrivals corresponding to the Bellcore trace `pAug.TL`

λ	K=250			K=500			K=1000		
	Sim.	Ana.	o/p queue	Sim.	Ana.	o/p queue	Sim.	Ana.	o/p queue
0.10	0.0544	0.0502	0.0506	0.0504	0.0484	0.0488	0.0439	0.0450	0.0454
0.15	0.0631	0.0541	0.0549	0.0513	0.0523	0.0531	0.0510	0.0489	0.0498
0.20	0.0681	0.0581	0.0591	0.0603	0.0563	0.0573	0.0546	0.0530	0.0540

Table 9
Loss probabilities in a 8 × 8 switch with a parallelism of 4 and self similar arrivals corresponding to the Bellcore trace `pAug.TL`

5 Effect of Output Hotspots

We now analyze the effect of a hotspot on output port h, $1 \leq h \leq N$. The hotspot is characterized by a higher arrival rate to that output port. For example, we can use,

$$\begin{aligned} p_{ij} &= \begin{cases} \beta & \text{for } j \neq h \\ \gamma\beta & \text{for } j = h,\ \gamma > 1 \end{cases} \\ \sum_{j=1}^{N} p_{ij} &= 1 \qquad\qquad \text{for all } i \end{aligned} \tag{26}$$

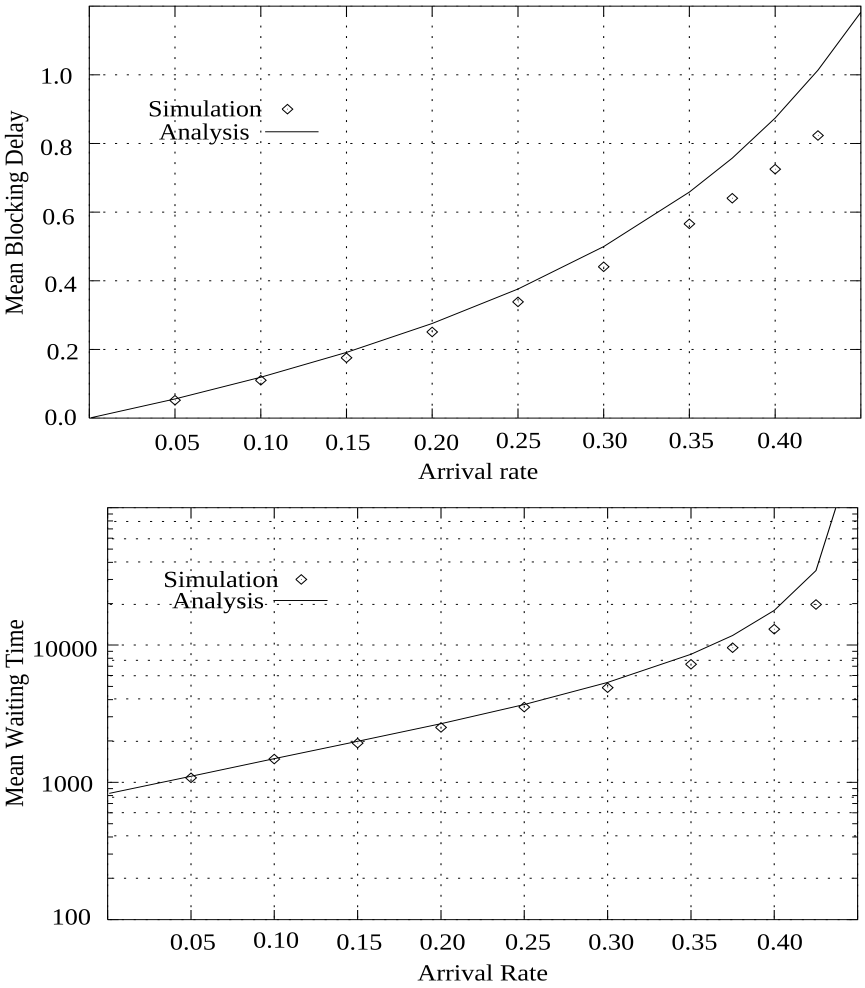


Fig. 11. Effect of an output hotspot on mean blocking and total delay for a 16 × 16 switch. The input process has the same characteristics as that of the Bellcore trace `pAug.TL`.

It is easy to see that as γ increases, the contention for the hotspot output port h increases and hence the blocking delay for these packets at the head of their input queues increases. The increased blocking delay increases the "input service time" and hence the total delay of all the packets. The switch can be analyzed using the methods that we outline in the previous sections. In this section we just report the numerical results for the case of infinite input buffers and self similar arrivals.

In Figure 11 we show the effect of this hotspot for $\gamma = 2$. As is evident from the figures, there is a marked rise in the average delays in the presence of hotspots and a considerable reduction in the maximum achievable throughput. A more detailed analysis can be carried using the analytical techniques that we have outlined earlier.

In our analytical model, asymmetry in the correlation or the Hurst parameter of the traffic at the input ports does not affect the delay performance of the other ports as long as the arrival rate remains constant. This is because the "service time" for a port depends on the blocking delay and the only factor affecting the blocking delay at the ports are the arrival rates into the virtual queues of the outputs. Thus the "service times" at all the ports in the presence of parameter asymmetries is the same. Hence, if the arrival rates are the same, differences in H, ρ and n do not have any effect on the "service times" of the other ports. However, the total delay at the ports will depend on the traffic characteristics at that input port.

6 Effect of ρ and H on the Delay Throughput Performance

Recall that the parameters in characterizing the input process are H the Hurst parameter, ρ the correlation a lag 1 and n the number of time scales over which burstiness occurs. In addition there are the routing probabilities and p_{ij} that can generate hotspots on some outputs. In this section we examine the effect of asymmetries in these parameters across the inputs on the throughput delay characteristics for the input queued switch.

Now let us consider the effect of the correlation structure of the arrival process at each input on the delay throughput characteristics. Figure 12 shows the effect of variation of the correlation on delay characteristics. The three graphs correspond to the case when the input processes have the same Hurst parameter ($H = 0.82$) and arrival rate but correlations at lag one of 0.532, 0.582 and 0.632. Each input port of the switch is fed with traces having the same parameters. Observe that the delay decreases substantially with lower correlations. This is due to the reduced probability of successive time units having packet arrivals and thus reducing the queuing at the inputs.

Finally we study the effect of variation in the Hurst Parameter. As in the previous case, we vary the Hurst parameter of the input streams keeping all other parameters constant. The delay throughput characteristics for the cases when the input steams at each port have Hurst parameters of 0.77, 0.82 and 0.87 for a correlation at lag one of 0.582 are shown in Figure 13. As before, each input port is fed with traces having the same statistical properties. Note that the delays decrease significantly with even slight reduction in the Hurst parameter. This can be explained by considering the fact that a lower H reduces the long range dependence and the burstiness thereby reducing the queue buildups at the inputs.

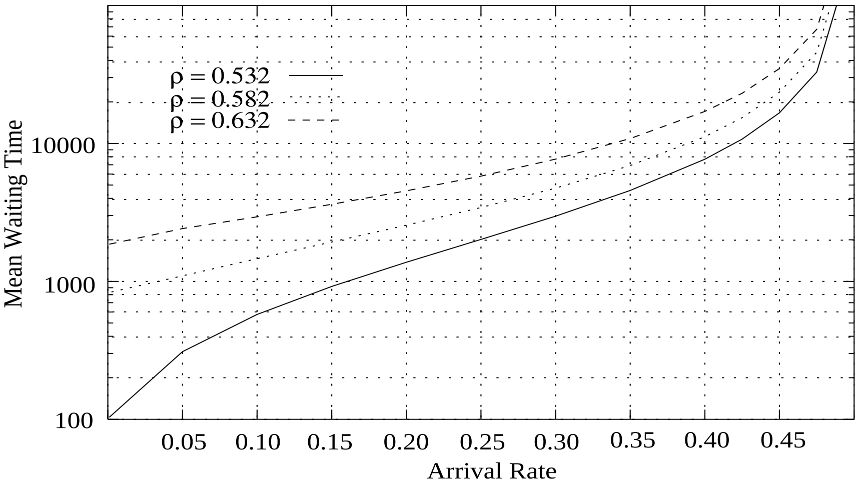


Fig. 12. The effect of variation in the correlation structure of the input traffic. We use $H = 0.82$ in the above results and the switch size is 16×16.

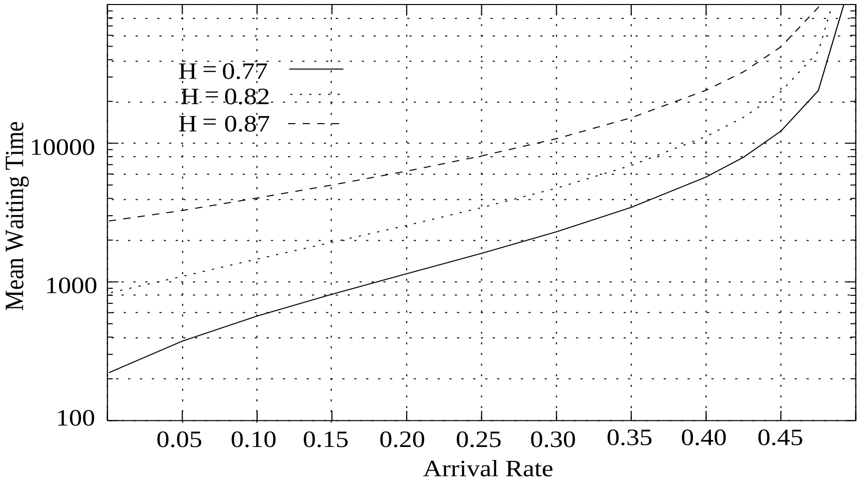


Fig. 13. The effect of variation in the Hurst parameter of the input traffic on different ports. We use $\rho = 0.582$ in the above results and the switch size is 16×16.

7 Conclusion

In this paper, we have presented a generalized analytical model for an input queued, variable length packet switch. We present models for both Poisson as well as self-similar arrivals and our model can be easily extended for any arrival process with a well defined queueing theory. Also, we have presented the analysis for switches with both infinite as well as finite input buffers. Our model can address random order (ROS) of service and priority models. For Poisson arrivals, it can be shown that the first moment of the waiting time for ROS would be the same as FIFO scheduling and the second moment $E(W_{fifo}^2)/(1 - \lambda * E(X_i))$ where $E(W_{fifo}^2)$ is the mean waiting time given by the first term of Eqn 8 and $E(X_i)$ is the expected service time obtained from Eqn 7. We could also easily extend the model to consider priorities in the input queue.

In [12] it was conjectured that FCFS service in the virtual output queue gives the least average delay. Our analysis easily confirms this because FCFS service has the least variance and this is the variance of the "service time" of the input queue which is an M/G/1 queue. It is well known that for an M/G/1 queue the variance of the service time, in addition to the mean, contributes to the average delay. Also, from our models it is clear that the conjecture in [12] that the performance of an $M \times N$ switch is symmetric in M and N is not true.

From the throughput-delay characteristics of Figures 5, 12 and 13 we see that capturing all the statistical properties of the arrival processes is essential to characterizing the switch performance. Another important result to note is that operation in continuous time limits the maximum achievable throughput to 0.5, though, with a parallelism of 4, the achievable throughput can be increased to more than 99%. Severe performance degradation takes place in the presence of hotspots, which can reduce the maximum throughput by 15% in a 16 × 16 switch. Also, Figures 12 and 13 highlight the large variations in the delay characteristics with changes in the correlation structure and the Hurst parameter. Lower Hurst parameters and correlation values reduce the burstiness of the arrival streams and reduces the queuing effects at the inputs and can give significantly lower delays at low loads. Thus the correlation structure and the Hurst parameter of the arrival processes are of considerable importance in determining the overall switch performance.

We now consider the implications of our results on the design of packet switches for variable length packets. We noted earlier that variable length packet can be switched using cell switches albeit with additional circuits for making cells from packets at the inputs and packets from cells at the outputs. However, such switches require a switching speed greater than the line rate to handle the increased load resulting from the padding of packets that don't fill an integral number of cells. To get an idea of the additional overhead involved in the fragmentation process due to the addition of header and padding we performed the following experiment. Let p_i be the number of bytes in packet i. Consider a switch that handles fixed length packets of size s. A space division switch will need to use its own header which will be local to the switch. Let h be the number of bytes in the switch header. The number of bytes transferred by the fixed length packet switch for packet i, b_i, will be $b_i = \lceil p_i/s \rceil * (s + h)$ and $p_i - b_i$ is the excess bytes per packet. We used the Ethernet and WAN packet traces that were used for the experiments reported in [18] and have obtained the mean and standard deviation of $p_i - b_i$ as a function of s for these packet traces. These are shown in Figure 14. Observe that even for the optimal choice of the cell size, we have an average overhead of 40-50 bytes per packet.

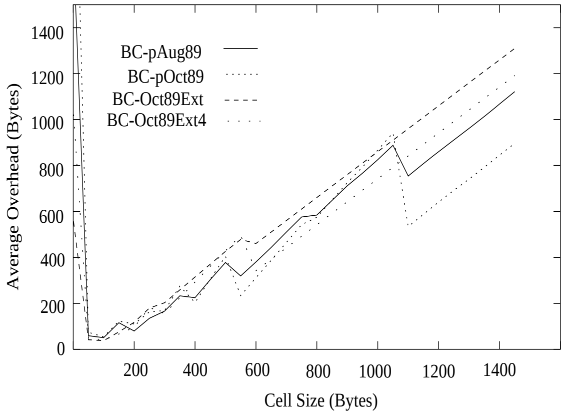


Fig. 14. The mean overhead incurred due to fragmentation for various cell sizes and a header length of 4 bytes. The traces used correspond to the Ethernet and WAN traces reported in [18].

Thus the additional complexities for variable length packet switches using cell switches leads us to believe that for variable length packet switches with non slotted arrivals, a FIFO-CIOQ packet switches with parallelism or speedup will be architecturally simpler than the VOQ-CIOQ cell switch. Further, our analysis suggests that from throughput delay performance perspective, for variable length packet switches, the FIFO-CIOQ with parallelism and/or speedup will have comparable performance to the time slotted, fixed packet length VOQ-CIOQ switches. The implementation tradeoff is in the addition of switching planes rather than a complex centralized scheduler that has to collect information about the N^2 queues and of course the “depacketization” and packetization circuit. Recall that a speedup of greater than two is required by the VOQ-CIOQ switch to have 100% throughput for arbitrary arrival processes. Also, although arbitrary QoS can be supported by the VOQ-CIOQ, the QoS scheduler has to be centralized and to be feasible in the Tbps region, the scheduling for VOQ-CIOQ switches will have to be done “on-chip” and propagation delays for going off chip with current technology will make it infeasible to implement VOQ-CIOQ switches at such speeds. Instead, a FIFO-CIOQ switch with speedup and parallelism can minimize the delay in the input queue and the output QoS scheduler can be used to provide the necessary QoS without requiring to have information about N^2 queues.

Acknowledgements

We would like to thank Prof. A. Baiocchi of University of Rome, Italy for giving us the code for analysizing a MAP/G/1/K queue and the reviewers for their comments.

A Delay Moments in an MMPP/G/1 Queue

The mean and second moment of the packet delay at input i, $\overline{D_i}$ and $\overline{D_i^2}$ respectively, are given by [11]

$$D_i = \frac{1}{\overline{X_i}\lambda_i}\left(w_v - \frac{1}{2}\overline{X_i^2}\lambda_i\right)$$

$$D_i^2 = \frac{1}{\overline{X_i}\lambda_i}\left(w_v^{(2)} - \frac{\overline{X_i^3}\lambda_i}{3} - D_i\overline{X_i^2}\lambda_i\right) \tag{A.1}$$

where

$$w_v = \frac{1}{2(1-\overline{X_i}\lambda_i)}\Big[2\overline{X_i}\lambda_i + \overline{X_i^2}\lambda_i - 2\overline{X_i}((1-\overline{X_i}\lambda_i)\mathbf{g}_i \\ + \overline{X_i}\mathbf{\Phi}_i\mathbf{R}_i)(\mathbf{Q}_i + \mathbf{e}_i\mathbf{\Phi}_i)^{-1}\mathbf{r}_i\Big]$$

$$w_v^{(2)} = \frac{1}{3(1-\overline{X_i}\lambda_i)}\Big[3\overline{X_i}(2\mathbf{W}_i'(0)(\overline{X_i}\mathbf{R}_i - \mathbf{I}) - \overline{X_i^2}\mathbf{\Phi}_i \\ \mathbf{R}_i)\overline{X_i^2}\mathbf{\Phi}_i\mathbf{R}_i)(\mathbf{Q}_i + \mathbf{e}_i\mathbf{\Phi}_i)^{-1}\mathbf{r}_i - 3\overline{X_i^2}\mathbf{W}_i'(0)h + \\ \overline{X_i^3}\lambda_i\Big]$$

with

$$\mathbf{W}_i'(0) = (\overline{X_i}\mathbf{\Phi}_i\mathbf{R}_i + (1-\overline{X_i}\lambda_i)\mathbf{g}_i)(\mathbf{Q}_i + \mathbf{e}_i\mathbf{\Phi}_i)^{-1} \\ -\mathbf{\Phi}_i(1+w_v)$$

and g_i representing the steady state probability vector of the matrix G_i, the transition rate matrix of the embedded Markov chain at departure epochs with k packets in the queue and the MMPP arrival process in state j. We now present the procedure for calculating the matrix G and a general algorithm to calculate the first and second moments of the delay in an MMPP/G/1 queue [11].

A.1 Computation of G_i for an m-state MMPP

Initial Step : Define

$$G_i^0 = 0 \qquad H_i^{0,k} = I \quad \text{for } k = 0, 1, 2, \cdots$$

$$\Theta = \max_j ((R_i - Q_i)_{jj})$$

$$\gamma_n = \int_0^\infty e^{-\Theta x} \frac{(\Theta x)^n}{n!} dH(x) \quad \text{for } n = 0, 1, \cdots, n^*$$

where n^* is chosen such that $\sum_{k=1}^{n^*} \gamma_k > 1 - \epsilon_1$, $\epsilon_1 \ll 1$.

Recursion : For $k = 0, 1, 2, \cdots$, do

$$H_i^{n+1,k} = \left[I + \frac{1}{\Theta}(Q_i - R_i + R_i G_i^k) \right] H_i^{n,k}$$

$$G_i^{k+1} = \sum_{n=0}^{n^*} \gamma_n H_i^{n,k}$$

Stopping Criterion :

$$|| G_i^{k-1} - G_i^k || < \epsilon_2 \ll 1$$

Set $G_i = G_i^{k+1}$.

A.2 Computation of γ_n

The γ_n for Erlang-k and exponential service times are given by

1. *Erlang-k service*

$$\begin{aligned} \gamma_n &= \int_o^\infty e^{-\Theta x} \frac{(\Theta x)^n}{n!} \mu^k \frac{x^{k-1}}{(k-1)!} e^{-\mu x} dx \\ &= \frac{(n+k-1)!}{n!(k-1)!} \frac{\mu^k \Theta^n}{(\Theta + \mu)^{n+k}} \end{aligned}$$

2. *Exponential service*

$$\begin{aligned} \gamma_n &= \int_o^\infty e^{-\Theta x} \frac{(\Theta x)^n}{n!} \mu e^{-\mu x} dx \\ &= \frac{\mu \Theta^n}{\Theta + \mu)^{n+1}} \end{aligned}$$

The γ_n for the service time distribution which is the summation of the phase-type distribution with Erlang-k service times and an exponential evacuation time is given by the weighted sum of the individual γ_n values. The weights are the probabilities of encountering each of the individual distributions, the $\pi_{ij}(k)$s.

A.3 The MMPP/G/1 algorithm

Step 1. Compute the matrix G for the given input port.
Step 2. Compute the steady state vector g which satisfies

$$g_i G_i = g_i \qquad g_i \epsilon = 1$$

Step 3. Compute the moments of the waiting time using Eqn A.1.

References

[1] A. T. Andersen and B. F. Nielsen, A Markovian approach for modeling packet traffic with long-range dependence, *IEEE Journal on Selected Areas in Communications* **16** (1998) 719-732.

[2] T. E. Anderson, S. S. Owicki, J. B. Saxe and C. .P Thacker, High speed switch scheduling for local area networks, *ACM Transactions on Computer Systems* **11** (1993) 319–352.

[3] A. Baiocchi and N. Bléfari-Melazzi, Steady-state analysis of the MMPP/G/1/K queue, *IEEE Transactions on Communications* **41** (1993) 531-534.

[4] A. Baiocchi and N. Bléfari-Melazzi, Analysis of the loss probability of the MAP/G/1/K queue Part II: Approximations and numerical results, *Stochastic Models* **10** (1994) 895-925.

[5] J. Chao, Saturn: A terabit packet switch using dual round robin, *IEEE Communications Magazine* **38** (2000) 78-84.

[6] K. J. Christensen, Design and evaluation of a parallel polled virtual output queued switch," *Proceedings IEEE ICC*, (Helsinki, Finland, 2001) 112-116.

[7] S-T. Chuang, A. Goel, N. McKeown and B. Prabhakar, Matching output queueing with a combined input/output-queued switch, *IEEE Journal on Selected Areas in Communications* **17** (1999) 1030–1039.

[8] J. G. Dai and B. Prabhakar, The throughput of data switches with and without speedup, *Proceedings of IEEE INFOCOM* (Tel Aviv, Israel, 2000) 556-564.

[9] A. Erramilli, R. P. Singh and P. Pruthi, An application of deterministic chaotic maps to model packet traffic, *Queuing Systems* **20** (1995) 171-206.

[10] A. Erramilli, O. Narayan and W. Wilinger, Experimental queuing analysis with LRD packet traffic, *IEEE/ACM Transactions on Networking* **4** (1996) 171-206.

[11] W. Fischer and K. Meier-Hellstern, The Markov-modulated Poisson process (MMPP) cookbook, *Performance Evaluation* **18** (1992) 149-171.

[12] S. W. Fuhrman, Performance of a packet switch with a crossbar architecture, *IEEE Transactions on Communications* **41** (1993) 486-491.

[13] E. Gelenbe and J. Pujolle, *Introduction to Queueing Networks* (John Wiley and Sons, 1987).

[14] M. Grossglauser and J. -C. Bolot, On the relevance of long-range dependence in network traffic, *IEEE/ACM Transactions on Networking* **7** (1999) 629-640.

[15] M. J. Karol, M. G. Hluchyj and S. P. Morgan, Input versus output queuing on a space-division packet switch, *IEEE Transactions on Communications* **35** (1987) 1347-1356.

[16] P. Krishna, N. S. Patel A. Charny and R. J. Simcoe, On the speedup required for work conserving crossbar switches, *IEEE Journal on Selected Areas in Communications* **17** (1999) 1057-1066.

[17] A. Kumar and L. Jacob, Stability condition for the input buffers of an input queueing ATM cell switch, *Unpublished Paper* (1996).

[18] W. E. Leland, M. S. Taqqu, W. Willinger and D. V. Wilson, On the self-similar nature of Ethernet traffic (extended version), *IEEE/ACM Transactions on Networking* **2** (1994) 1-15.

[19] S. Q. Li, Performance of a non-blocking space division packet switch with correlated input traffic, *Proceedings of IEEE GLOBECOM* (Dallas, Texas, 1989) 1754-1763.

[20] D. Manjunath and B. Sikdar, Analysis of input queued switches for variable length packets, *RPI ECSE Networks Laboratory Technical Report, ECSE-NET-2001-1* (Troy, NY, 2001).

[21] N. McKeown, *Scheduling Algorithms for Input-Queued Cell Switches*, (Ph. D. Thesis, University of California, Berkeley, 1995).

[22] N. McKeown, V. Anantharam and J. Walrand, Achieving 100% throughput in an input queued switch, *Proceedings of IEEE INFOCOM* (San Francisco, California, 1996) 296-302.

[23] N. McKeown and A. Mekkittikul, A starvation free algorithm for achieving 100% throughput in an input queued switch, *Proceedings of IEEE ICCCN* (Washington, DC, 1996) 226-231.

[24] N. McKeown and A. Mekkittikul, A pratical scheduling algorithm to achieve 100% throughput in input-queued switches, *Proceedings of IEEE INFOCOM* (San Francisco, California, 1998) 792-799.

[25] N. McKeown, iSLIP: A scheduling algorithm for input output queued switches, *IEEE/ACM Transactions on Networking* **7** (1999) 188-201.

[26] N. McKeown, M. Izzard, A. Mekkittikul, W. Ellersick and M. Horowitz, Tiny Tera: A packet switch core, *IEEE Micro* **17** (1997), 26-33.

[27] G. Nong and M. Hamdi, On the provision of quality of service guarantees for input queued switches, *IEEE Communications Magazine* **38** (2000) 62-69.

[28] Y. Oie, M. Murata, K. Kubota and H. Miyahara, Effect of speedup in nonblocking packet switch, *Proceedings of IEEE ICC* (Boston, Massachussets, 1989) 410-414.

[29] J. H. Patel, Performance of processor-memory interconnections for multiprocessors, *IEEE Transactions on Computers* **30** (1981) 771-780.

[30] V. Paxson and S. Floyd, Wide area traffic : The failure of Poisson modeling, *IEEE/ACM Transactions on Networking* **3** (1995) 226-244.

[31] H. Perros, *Queuing Networks with Blocking* (Oxford University Press, 1994).

[32] K. Shiomoto, M. Uga, M. Omotani, S. Shimizu and T. Chimaru, Scalable multi-QoS IP+ATM switch router architecture, *IEEE Communications Magazine* **38** (2000) 86-92.

[33] B. Sikdar and K. S. Vastola, On the convergence of Markovian and Fractional ARIMA processes with long-range dependence to fractional Brownian motion, *Proceedings of 34th Conference on Information Sciences and Systems* (Princeton, NJ, 2000) TP2(7-12).

[34] M. A. Stephens, EDF statistics for goodness of fit and some comparisons, *Journal of the American Statistical Association* **69** (1974) 730-737.

[35] K. Yoshigoe and K. Christensen, A parallel polled virtual output queued switch with a buffered crossbar, *Proceedings of IEEE HPSR* (Dallas, Texas, 2001) 271-275.