Analysis of Shared Buffer Multistage Networks with Hot Spot

Mahmoud Saleh
Dept. of Comp. Science and Comp. Engg.
LaTrobe University
Melbourne, 3083, Australia
Email: *saleh@cs.latrobe.edu.au*

Mohammed Atiquzzaman
Dept. of Elec. & Computer Systems Engg.
Monash University, Clayton,
Melbourne 3168, Australia
Email: *atiq@eng.monash.edu.au*

Abstract

Multistage Interconnection Networks based on shared buffering are known to have better performance and buffer utilization than input or output buffered switches. Shared buffer switches do not suffer from head of line blocking which is a common problem in simple input buffering. Shared buffer switches have previously been studied under uniform and unbalanced traffic patterns. However, due to the complexity of the model, the performance of such a network, in the presence of a single hot spot, has not been fully explored. A hot spot arises when one of the outputs of the network becomes very popular. In this paper, we develop a model for a multistage interconnection network constructed from shared buffer switching elements and operating under a hot spot traffic pattern. The model is validated by comparison with simulation results. The model is used to study the network performance in terms of the throughput, packet delay, packet loss probability and the optimal buffer utilization. Numerical results show that, in the presence of hot spot traffic, shared buffer switches degrade more significantly than switches with dedicated input and/or output buffers.

1 Introduction

Multistage Interconnection Networks (MINs) have been widely studied as efficient interconnection structures for parallel computer systems. Also, in recent years, MINs have received increasing attention as the switching fabrics for broadband integrated services digital network (B-ISDN) and transport systems based on asynchronous transfer mode (ATM). An ATM network transfers all information in fixed length packets called *cells*, and is characterized by simplified protocols, high speed links, and high capacity switching nodes. MINs are particularly useful for ATM switches for the features they offer, such as modularity and decentralized routability.

MINs have also been found to be suitable for interconnecting a large number of processors and memories in a multiprocessor system. Consequently, the performance of MINs has been widely studied for multiprocessors systems [1, 2, 3, 4, 5, 6, 7] and communications networks [8, 9, 10, 12, 13].

A MIN consists of a number of stages of small switching elements (SE) which are interconnected by a permutation function. A blocking type MIN suffers from packets contending for the same outlet within the switch which results in a loss in the performance of the switch. Delta, Omega, and Banyan networks are examples of blocking types of MINs. Performance of such networks can be increased by using a sorting network at the input of the network, or by having multiple paths between input/output pairs, or by using buffers to store the conflicting packets. Multiple path networks need additional control mechanisms to manage multiple submission of packets to different paths. Internally buffered networks employ buffers at the SEs inside the network. The packets losing contentions at the SEs are stored in the buffers in the SEs. The location of buffers in an SE is crucial in the throughput, delay, and cost of the network. FIFO buffers placed at the inlets of the SEs suffer from head of line (HOL) blocking and result in reduced throughput. Input queues with bypass have been suggested to reduce the effect of HOL contention. Buffers may be placed at the outlets of the SEs, and the packets destined to a particular outlet of an SE are queued at the corresponding buffer. In an output buffered network, a $d \times d$ SE requires reduced buffer access time and internal speedup which is d times the switching speed of an input-buffered SE.

Due to the use of dedicated buffers for the inputs or outputs, the networks constructed from input or output buffered SEs have low buffer utilization for most unbalanced traffics. Shared buffers may be used in the SEs to increase the buffer utilization and the performance of the network. Buffers in a shared buffer SE are shared by the inlets and the outlets of the SE such that a packet coming to an inlet may be placed into any available shared buffer in the SE, and a packet in a buffer can be forwarded to any of the outlets. An SE employing shared buffers does not suffer from HOL blocking. In addition, unlike output buffering, buffer resources in shared buffering are allocated to the outputs which most need them, and not dedicated to a particular output regardless of its needs. Consequently, MINs constructed from shared buffer SEs have higher throughput, lower delay and better buffer utilization than networks constructed from input or output buffered SEs. Moreover, given the same amount of buffer, the shared buffer is the best choice in terms of packet loss rate [9, 14, 15]. Since one of the important performance criteria for ATM networks is packet loss rate, a shared buffer architecture is very suitable for implementing ATM networks in B-ISDN.

Performance evaluation studies may be accomplished

0-7803-2018-2/95/$4.00

by simulation or analytical modeling. Although the simulation enables one to closely study the behavior of a network, using simulation to estimate the probability of rare events and their effect on performance is problematic, because vast computational resources may be required to generate a sufficient number of events from which statistical estimates may be formed with adequate statistical confidence [11]. In the analytical modeling, on the other hand, the results are obtained much faster with no special attention to calculation of very small probabilities. However, an analytical model often needs to be simplified by restrictive assumptions in order to be mathematically tractable.

Turner [12] developed a model for a Delta [17] network with shared buffer SEs under uniform traffic distribution. His model assumes independence between buffer slots, and uses some flow control mechanism to avoid packet loss inside the network. This model was extended by Monterosso [10] and Bianchi [19] for more accurate models. A model for a network using shared buffer SEs, operating under a uniform traffic pattern and global flow control policy, has been reported in [13]. Gianatti and Pattavina [20] studied shared buffer networks with nonuniform traffic patterns. However, in this model, the outputs of the MIN are divided such that a group of outputs are hot and the rest are cold. The number of SEs in the hot group is determined by $\log_d N$, where N is the network size, and d is the size of an SE. For example, for $N = 64$, and $d = 2$, they consider 32 hot, and 32 cold outputs. Hence, the model is not suitable for studying networks with a single hot output, where one of the network outputs becomes more popular than the others. A hot spot can be observed in a distributed operating system, when tasks demand more frequent access to some system resource, or, in a loosely coupled system, when a file server or a printer is accessed more likely than the other resources of the system. In the ATM networks a hot spot situation may happen when some particular trunk receives more traffic than the rest of the outputs. Previous simulation studies have shown the detrimental effect of hot spots on the performance of shared buffer networks [18].

Most of the above models use local flow control to control packet movement between stages. In local flow control, a packet can be forwarded to the next stage depending on its state at the beginning of a cycle, whereas in global flow control the simultaneous operations of forwarding and receiving packets during a cycle are allowed. Therefore, global flow control results in a higher throughput and better buffer utilization than local flow control.

The *aim* of this paper is to study the performance of a multistage Delta network with global flow control and operating under a hot spot traffic pattern. Our *objectives* are:

1. to develop a model for Delta networks using *shared buffer* SEs and operating under *single hot spot* traffic pattern;
2. to determine the effect of hot spot traffic on throughput, delay and packet loss probability;
3. to study the effect of a hot spot on the *buffer utilization* of different SEs.

This paper is organized as follows. In Section 2, we describe the modeling assumptions and single hot spot model. Construction of the corresponding simulator, and additional considerations are explained in Section 3. In Section 4, we examine our model with some numerical examples, and compare the results with the simulation. Concluding remarks and further possible work are given in Section 5.

2 Shared Buffer Delta Network

In this section, we develop a model for a shared buffer Delta network operating under a hot spot traffic pattern. A Delta-d network with N inlets and N outlets consists of k stages of $d \times d$ SEs such that $N = d^k$. Starting from the stage where packets are offered to the network, we number the stages from 1 to k. In a Delta$-d$ network, there exists only one path between each input and output of the network, and each stage of the network consists of N/d SEs. At stage i (of a single hot spot Delta network), there are i types of SEs which carry different mixtures of hot and cold traffic [4]. A hot spot Delta-2 network is illustrated in Fig. 1 where the hot links and SEs carrying hot traffic are shown by thick solid lines.

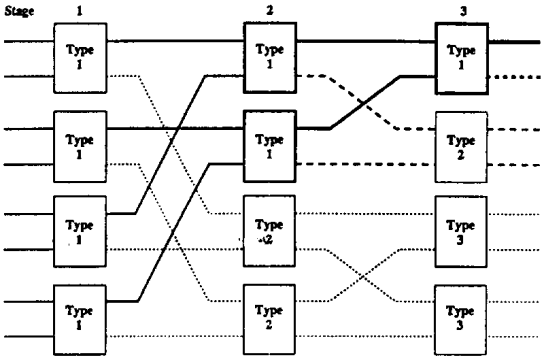

Figure 1: An 8×8 Delta-2 MIN with single hot spot.

2.1 Assumptions

The following assumptions are made regarding the network and its operation:

- Each SE is of size $d \times d$ and contains B buffers which are shared by the d inlets and d outlets of the SE.
- The network operates *synchronously*, i.e. packets are submitted to the network at the beginning of the cycles.
- *Destination tag* is used to route a packet. A routing conflict inside the network is resolved *randomly*, i.e. if two or more packets are destined to the same output, one is chosen at random.
- The *state* of an SE whose buffers contain exactly $s = h + c$ packets is represented by a pair (h, c) where h is the number of packets destined to the hot outlet of the SE and c is the number of packets

destined to the other $d-1$ cold outlets of the SE. We label the hot SEs as *type 1* SEs.

- The arrival of packets at each input of the network is a *Bernoulli* process, i.e., the probability that a packet arrives during a cycle (ρ) is constant, and arrivals are independent of each other.

- The probability of a packet arriving at a network input and being destined to the *hot* output (p_h) or to any single one of the $N-1$ *cold* outputs (p_c) is given by:

$$p_h = \rho\left(f_h + \frac{1-f_h}{N}\right), \quad p_c = \rho\left(\frac{1-f_h}{N}\right),$$
$$p_h + (N-1)p_c = \rho. \qquad (1)$$

  where f_h is the fraction of the hot traffic.

- Since there exists only one path between an input and an output of the network, a hot output can be reached through only one of the outlets of an SE (hot outlet). An incoming packet loads the other $d-1$ non-hot outlets of an SE with equal probability.

- A *backpressure* mechanism with global flow control ensures that no packet is lost inside the network. Thus, a packet leaves an SE if there is a space for it in the next stage SE, or if a space becomes available during the same cycle. An acknowledgment policy is used to advise the receipt of a packet in the next stage SE. Unacknowledged packets contend with other packets in subsequent cycles.

- There is no *blocking* at an output of the network, i.e., an output can always accept a packet.

For the purpose of analysis, we assume that the process of forwarding and accepting packets in each SE is accomplished in two phases [13]. In the *forward* phase, depending on the state of the SE and its downstream SEs, a number of packets may leave the SE, and the switch goes to an *intermediate* state. During the *receive* phase, the packets offered from upstream SEs are placed in the buffers, the corresponding acknowledgments are sent to the upstream SEs, and the SE goes to the final state. If the number of arriving packets is greater than the number of available buffers in the SE, a number of packets equal to the number of available spaces are selected randomly. The possible transitions of states in an SE for $d=2$ and $B=2$ are illustrated in Fig. 2.

2.2 Analysis of the MIN

We model each SE by a Markov chain representing the distribution of the hot and cold packets stored in the B buffers of the SE. An SE is of type i if it is fed by a type $i-1$ SE in the previous stage (Figure 1). It has been shown in [4] that stage i will have i different types of SEs and $i+1$ different traffic rates at its outlets.

The following notations will be used in the model.

$SE_{i,r}$: an SE of type r at stage i.

$\pi_{i,r,t}(h1,c1)$: probability that $SE_{i,r}$ is in state $(h1,c1)$ at the beginning of cycle t.

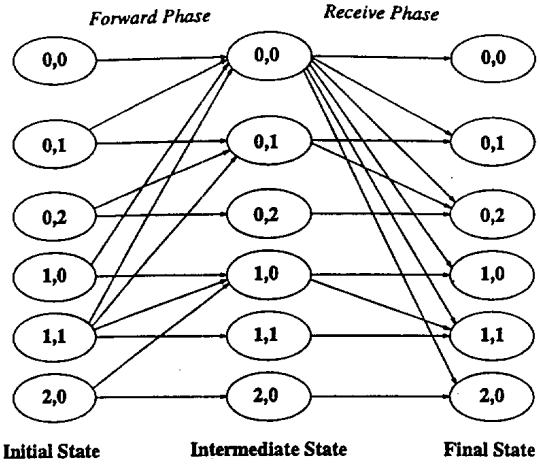

Figure 2: State diagram of a two phase network operation in an SE with $d=2$, and $B=2$. Every state is denoted by a pair (h,c) where h is the number of packets destined to the hot outlet of the SE and c is the number of packets destined to the other $d-1$ cold outlets of the SE.

$\tau_{i,r,t}(h1,c1,h3,c3)$: probability that $SE_{i,r}$ is in state $(h3,c3)$ at the beginning of the receive phase, given that it was in state $(h1,c1)$, at the beginning of the forward phase of cycle t, where $0 \le h1-h3 \le 1$, and $0 \le c1-c3 \le d-1$.

$\sigma_{i,r,t}(h3,c3,h2,c2)$: probability that $SE_{i,r}$ is in state $(h2,c2)$ at the end of the receive phase of cycle t, given that it was in state $(h3,c3)$ at the beginning of the receive phase of the same cycle, where $h3 \le h2$, $c3 \le c2$, and $h3+c3 \le h2+c2+d$.

$\tilde{\pi}_{i,r,t}(h3,c3)$: probability that $SE_{i,r}$ is in state $(h3,c3)$ at the beginning of the receive phase of cycle t.

$a_{i,r,t}$: probability that a packet is ready to enter $SE_{i,r}$ at cycle t.

$b_{i,r,t,x}$: probability that, at cycle t, a successor of $SE_{i,r}$ provides an acknowledgment to type x outlet of the SE, given that a packet was submitted to the successor through outlet x during the same cycle. x is of either a *hot* or *cold* outlet.

$Y_d(r,c)$: probability that c packets in an SE are destined to r distinct outlets of the SE from a total of d outlets under consideration.

$u_{i,r,j}$: probability that a packet in $SE_{i,r}$ is destined to its j^{th} outlet, where $1 \le j \le d$.

Our modeling approach is based on the iterative solution [6] of a Markov chain system which characterizes the behavior of $SE_{i,r}$ for different i and r in a Delta network. In this approach, if there exists a solution to the system, starting from an arbitrary initial state the

iterative results converge to the steady state condition of the system.

In $SE_{i,r}$, the Markov chain system is described as:

$$\tilde{\pi}_{i,r,t}(h3,c3) = \sum_{h1=0}^{B}\sum_{c1=0}^{B-h1} \pi_{i,r,t}(h1,c1)\times \tau_{i,r,t}(h1,c1,h3,c3) \quad (2)$$

$$\pi_{i,r,t+1}(h2,c2) = \sum_{h3=0}^{B}\sum_{c3=0}^{B-h3} \tilde{\pi}_{i,r,t}(h3,c3)\times \sigma_{i,r,t}(h3,c3,h2,c2) \quad (3)$$

where B represents the total buffer space in an SE. Eq. (2) states that the probability of $SE_{i,r}$ being in the intermediate state $\tilde{\pi}_{i,r,t}(h3,c3)$ is determined by $SE_{i,r}$ being in the initial state $\pi_{i,r,t}(h1,c1)$, and the probability of the transition $\tau_{i,r,t}(h1,c1,h3,c3)$ taking place, summing over all possible $h1$ and $c1$. Similarly, in Eq. (3), being in state $\pi_{i,r,t+1}(h2,c2)$ requires $SE_{i,r}$ to be in the intermediate state $\tilde{\pi}_{i,r,t}(h3,c3)$, and the transition $\sigma_{i,r,t}(h3,c3,h2,c2)$ to take place, summing over all possible $h3$ and $c3$.

The complete set of states is obtained by calculating the steady state vector $\mathbf{\Pi}_{i,r}$ for every type r SE at all of the stages where:

$$\mathbf{\Pi}_{i,r} = [\pi_{i,r}(h,c)], h = 0,\cdots,B; c = 0,\cdots,B-h. \quad (4)$$

The corresponding measurements of throughput, packet loss, and delay can then be derived from the set and transition equations.

For the rest of this paper, we drop the time subscript t in time dependent functions, considering that the Markov chain system is in its equilibrium condition.

Transition $\tau_{i,r}(h1,c1,h3,c3)$ in Eq. (2) takes place if $h1-h3$ packets leave $SE_{i,r}$ from its hot outlet, and $c1-c3$ packets leave the SE from its $d-1$ cold outlets:

$$\tau_{i,r}(h1,c1,h3,c3) = \beta(\min(1,h1),h1-h3,b_{i,r,hot})\times \sum_{l=c1-c3}^{d-1} Y_{d-1}(l,c1)\beta(l,c1-c3,b_{i,r,cold}), \quad (5)$$

where d is the number of inlets and outlets of an SE, and β is the shorthand notation for a binomial distribution;

$$\beta(n,k,p) = \binom{n}{k}p^k(1-p)^{n-k}. \quad (6)$$

Since we assume that the cold traffic of an SE is distributed uniformly over all $d-1$ outlets of an SE, Y_d may be obtained by [19]:

$$Y_d = \binom{d}{c}\frac{\gamma(s-c,c)}{\gamma(s,d)}, \quad (7)$$

where

$$\gamma(s,d) = \binom{s+d-1}{s}. \quad (8)$$

Y_d is independent of SE type and stage, so it can be calculated once and used for subsequent calculations.

$b_{i,r,x}$, the probability that a packet sent through an outlet of type x of $SE_{i,r}$ is accepted by its next stage depends on the stage and type of the SE.

1. $i = k$,
   Since there is no blocking at the outputs of the network, the probability $b_{i,r,x}$ of acceptance of an offered packet at stage k is equal to 1.

2. $i < k$,
   An offered packet to a particular outlet of an SE is definitely accepted by its successor SE, if there are at least d buffers in the successor SE, or if the total number of packets that are offered to other $d-1$ inlets of the successor SE is less than the available buffers in that SE. Otherwise, only a fraction of packets are acknowledged:

$$b_{i,r,x} = \sum_{h3=0}^{B-d}\sum_{c3=0}^{B-d-h3} \tilde{\pi}_{i+1,s}(h3,c3) + \sum_{h3=0}^{L1}\sum_{c3=0}^{B} \tilde{\pi}_{i+1,s}(h3,c3)\times \left[\sum_{w_h=0}^{L2}\sum_{w_c=0}^{d-1} \mu(a_{i+1,s}u_{i+1,s,hot},d-1,w_h,w_c) + \sum_{w_h=0}^{d-1}\sum_{w_c=0}^{L3} \mu(a_{i+1,s}u_{i+1,s,hot},d-1,w_h,w_c)\frac{B-(h3+c3)}{w_h+w_c+1}\right], \quad (9)$$

in which $L1$, $L2$ and $L3$ are restrictions on the upper limits of the summations and are given by:

$$\begin{aligned} L1 &= B-d+1 \le h3+c3 \le B-1, \\ L2 &= w_h+w_c \le B-(h3+c3)-1, \\ L3 &= B-(h3+c3) \le w_h+w_c \le d-1. \end{aligned}$$

Subscript s used in Eq. (9) denotes the type of SE which should be considered at the next stage as follows:

$$s = \begin{cases} 1 & ,i=1,x=1 \\ r+1 & ,r>1 \vee (r=1 \wedge x \neq 1) \end{cases}; \quad (10)$$

and $\mu(a,u,d,h,c)$ is the multinomial distribution of h and c from a total of d:

$$\mu(a,u,d,h,c) = \frac{d!}{h!c!(d-h-c)!}\times (a.u)^h[a(1-u)]^c(1-a)^{d-h-c}. \quad (11)$$

$\sigma_{i,r}(h3,c3,h2,c2)$ is calculated based on the knowledge we can obtain from the final state $(h2,c2)$. If there is still some buffer space after the transition from state $(h3,c3)$ to $(h2,c2)$ takes place, it means that every packet which has been offered to the current SE has been accepted. If, however, after the transition, no buffer in the SE is empty $(h2+c2=B)$, there may have been some contention for a packet to enter the SE, and hence, only a fraction of offered packets may have been

accepted.

$$\sigma_{i,r}(h3,c3,h2,c2) = \begin{cases} \mu(a_{i,r},u_{i,r,hot},d,h2-h3,c2-c3) & ,h2+c2<B \\ \sum_{w_h=h2-h3}^{d-(c2-c3)} \sum_{w_c=c2-c3}^{d-w_h} \frac{\binom{w_h}{h2-h3}\binom{w_c}{c2-c3}}{\binom{w_h+w_c}{c2-c3+h2-h3}} \times & \\ \mu(a_{i,r},u_{i,r,hot},d,w_h,w_c) & ,h2+c2=B \end{cases} \quad (12)$$

$a_{i,r}$ of the first stage is simply ρ, the input load of the network. In stages other than the first, $a_{i,r}$ depends on the state of the predecessor SE:

$$a_{i,r} = \begin{cases} \rho & ,i=1 \\ 1-\sum_{j=0}^{B}\pi_{i-1,r}(0,j) & ,i>1,r=1 \\ \sum_{l=1}^{B}\sum_{j=0}^{B-l}\pi_{i-1,r-1}(j,l)\times & \\ \left(\left(1-\frac{\gamma(s,d-2)}{\gamma(s,d-1)}\right)\right) & ,i>1,r>1 \end{cases} \quad (13)$$

$u_{i,r,j}$, the probability that a packet in $SE_{i,r}$ is destined to j^{th} outlet of the SE, is determined by:

$$u_{i,r,j} = \frac{enum_{i,r,j}}{denom_{i,r}}. \quad (14)$$

For the last stage, the probability $enum_{k,r,j}$ that a packet is referencing a hot or cold output is simply calculated by:

$$enum_{k,r,j} = \begin{cases} p_h & ,r=1 \wedge j=1 \\ p_c & ,(r=1 \wedge j>1) \vee r>1 \end{cases}. \quad (15)$$

For $i<k$, $enum_{i,r,j}$ is calculated recursively:

$$enum_{i,r,j} = \begin{cases} \sum_{h=1}^{d} enum_{i+1,r,h} & ,r=1 \wedge j=1 \\ \sum_{h=1}^{d} enum_{i+1,r+1,h} & ,r=1 \wedge j>1 \\ enum_{i+1,r,j} \times d & ,r>1 \end{cases}. \quad (16)$$

The probability $denom_{i,r}$ that, for every stage, any output of the Delta network accessible from $SE_{i,r}$ is referenced by a packet inside that SE is given by:

$$denom_{i,r} = \begin{cases} p_h + (d-1)p_c \sum_{h=i}^{k} d^{k-h} & ,r=1 \\ p_c d^{k-i+1} & ,r>1 \end{cases}. \quad (17)$$

2.3 Performance Evaluation

In steady state condition of the network, the throughput, packet loss, and delay of various SE types can be computed. Throughput of the hot outlet of an SE is equal to the sum of all possible transitions from an initial state (h,c) to intermediate state $(h-1,c)$, since there is only one outlet (hot) through which hot packets can leave the SE:

$$\lambda_{i,r,hot} = \sum_{h1=1}^{B}\sum_{c1=0}^{B-h1}\pi_{i,r}(h1,c1)\times \sum_{c3=0}^{c1}\tau_{i,r}(h1,c1,h1-1,c3). \quad (18)$$

Similarly, we are able to calculate the throughput of each cold outlet of an SE by dividing the cumulative throughput of all cold outlets by $d-1$, based on the assumption that all cold outlets are equally likely:

$$\lambda_{i,r,cold} = \sum_{c1=1}^{B}\sum_{h1=0}^{B-c1}\pi_{i,r}(h1,c1)\times \sum_{h3=0}^{h1}\sum_{c3=\max(0,c1-d)}^{c1}\frac{c1-c3}{d-1}\tau_{i,r}(h1,c1,h3,c3). \quad (19)$$

Summing the hot and cold throughputs of an SE, we get the overall throughput of that SE:

$$\lambda_{i,r} = \lambda_{i,r,hot} + (d-1)\lambda_{i,r,cold}. \quad (20)$$

Finally, the throughput of stage i is given by:

$$\Lambda_i = d^{k-i}\left[\lambda_{i,1} + (d-1)\sum_{r=2}^{i}\lambda_{i,r}d^{r-2}\right]. \quad (21)$$

Eq. (21) applies to all of the stages including the first stage where $\sum$ becomes irrelevant.

Since there is no packet loss inside the network, the throughputs of all stages are the same. Thus, to calculate the packet loss probability η, we use the throughput of the first stage for convenience.

$$\eta = \frac{\rho N - \Lambda_1}{\rho N} = \frac{\rho - \Lambda_1/N}{\rho}, \quad (22)$$

where Λ_i/N is the throughput per link in stage i.

Delay of hot and cold outlets of an SE may be calculated using Little's formula for delay in which waiting time in a queue is equal to the average queue length divided by the arrival rate of the queue. In a shared buffer SE, each outlet has a logical queue whose length is equal to the number of packets which are currently passed through that outlet. Since we assume no packet loss inside the network, the input and output rates of an SE are the same. Thus, we can use throughput equations to calculate the delay. For the hot outlet we have:

$$w_{i,r,hot} = \frac{1}{\lambda_{i,r,hot}}\sum_{h=0}^{B}\sum_{c=0}^{B-h} h\pi_{i,r}(h,c), \quad (23)$$

where $\lambda_{i,r,hot}$ is the departure rate, and the summation comprises the average queue length. $w_{i,r,cold}$ is defined as the average waiting time of a packet in any one of the cold outlet logical queues:

$$w_{i,r,cold} = \frac{1}{\lambda_{i,r,cold}}\left(\sum_{h=0}^{B}\sum_{c=0}^{B-h} c\pi_{i,r}(h,c)\right)\frac{1}{d-1} = \frac{1}{(d-1)\lambda_{i,r,cold}}\sum_{h=0}^{B}\sum_{c=0}^{B-h} c\pi_{i,r}(h,c). \quad (24)$$

Since all $d-1$ cold outlets of $SE_{i,r}$ have equal throughputs, the average logical queue length for any cold outlet of the SE in Eq. (24) is the sum of the average equivalent logical queue for all $d-1$ outlets divided by $d-1$. The average waiting time in $SE_{i,r}$ is obtained by summing the average logical queues of all outlets divided by d:

$$w_{i,r,av} = \frac{w_{i,r,hot} + (d-1)w_{i,r,cold}}{d}. \quad (25)$$

The average delay at stage i is equal to the sum of all $w_{i,r,av}$ for $1 \le r \le i$, divided by the number of SEs in the stage (N/d):

$$\begin{aligned} w_i &= \frac{N}{d^i}\left[w_{i,1,av} + (d-1)\sum_{r=2}^{i} w_{i,r,av} d^{r-1}\right] \frac{1}{N/d} \\ &= \frac{1}{d^{i-1}}\left[w_{i,1,av} + (d-1)\sum_{r=2}^{i} w_{i,r,av} d^{r-1}\right] \end{aligned} \quad (26)$$

Finally, the average overall delay is obtained by summing the delays in different stages of the network:

$$W = \sum_{i=1}^{k} w_i, \quad (27)$$

where k is the number of stages in the network.

3 Simulation Study

We validated the model presented in Section 2 with a simulation study. The same assumptions as made for the analysis apply to the simulation of the network, and the following operations are carried out:

- At each cycle, a packet is generated with probability ρ (offered load to the network input). The generated packet is independent of the packets generated in previous cycles and at other input ports. Each packet consists of the following information:
  1. a source tag which denotes the input link at which the packet arrived,
  2. a destination tag denoting the output link to which the packet is destined, and
  3. the current cycle number, used for measurement of the packet delay in the network.
- Simulation results from the first several hundred cycles of the network operation are ignored to allow the network to reach a steady state condition. The simulation program is then allowed to run until the change in the average throughput between consecutive cycles becomes less than 10^{-6}.
- Conflict in the buffers for accessing a particular outlet as well as contention to seize a buffer space in the next stage is resolved using a random number generator with a different seed value from that of the packet generator.

The network operates as follows:

1. The packets at the last stage buffers are sent to the output links of the network, and the instantaneous throughput and delay are measured for every link.
2. For each SE at stages $k-1$ to 1:
   - The SE buffers are examined for packets passing the different outlets of the SE, copies of all packets passing different outlets are placed in the corresponding outlet lists (forming logical output queues), and the lists are sent to the corresponding inlets of the next stage.
   - If the number of available buffer spaces in the SE is less than the number of packets in the different lists at the inlets to the SE, a number of packets equal to the number of available spaces are chosen at random from the available lists. Packets which are not accepted stay in the buffers at the previous stage until they can be forwarded in the subsequent cycles.
3. A new set of packets are generated at the inputs of stage 1 with probability ρ and hot spot probability f_h, and packets are placed in the first stage buffers if there is any room. If a packet can not be placed in the first stage buffers, it is discarded, and the packet loss counter is incremented by one.

4 Numerical Results

Normalized throughput of a Delta network for $N = 256$, $d = 2$, and hot spot values 0 (uniform traffic) and 0.005 is illustrated in Fig. 3. In this figure, the proposed model is quite accurate when the input load is small. When the input load is more than 0.4, the model is still accurate when buffer size is small ($B = 2$), and the hot spot value is more than 0.05. The model is optimistic for larger buffer sizes ($B = 4$), or smaller hot spot values, since it assumes that output addresses of the packets are independent of each other; whereas in reality, a blocked packet which attempts a particular outlet of an SE in the current cycle will definitely attempt the same outlet in the subsequent cycles, too. The probability of a packet being blocked increases as the input load increases, and so does the discrepancy between the results from the model and the simulation.

The average delay W of the same network is shown in Fig. 4. The results from the proposed model are consistent with the simulation results for buffer sizes two and four. As in Fig. 3, the results from the model are close to those of the simulation for small input loads. The model predicts lower delay time due to the same fact that it ignores the effect of blocked packets which in practice cause higher delays in a network. In Fig. 5, the packet loss for uniform traffic and various hot spot values and buffer sizes is illustrated. The packet loss in a shared buffer MIN is very low when the B/d ratio is large, and the traffic is uniform. However, when a hot spot value is introduced into the network, the packet loss increases sharply as the input load increases. This is because under hot spot traffic, formation of tree saturation [21] directly affects the rate of packet acceptance in a buffer.

Though the throughput of the hot output of a network increases sharply when the hot spot value increases, the overall throughput of the network decreases due to the buffer monopolization effect [16] caused by the hot traffic. This situation has a greater impact on the overall

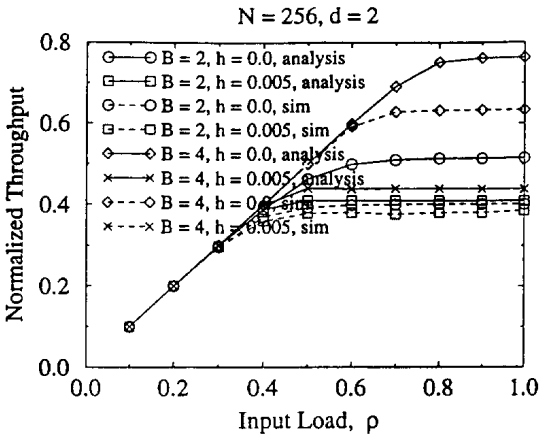

Figure 3: Normalized throughput versus ρ for $N = 256$, and $d = 2$.

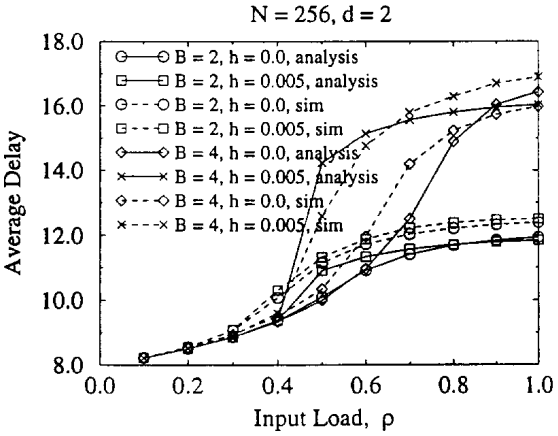

Figure 4: Average delay versus ρ for $N = 256$, and $d = 2$.

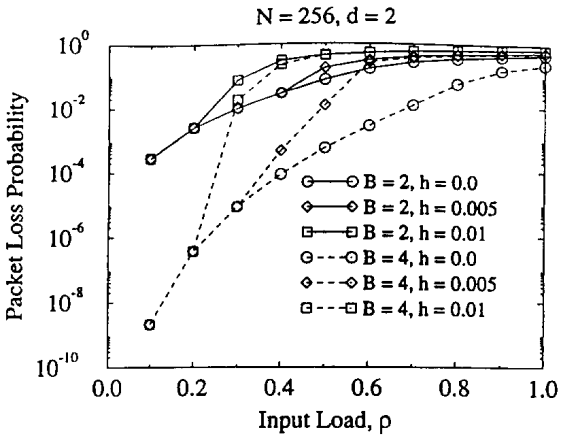

Figure 5: Packet loss versus ρ for $N = 256$, and $d = 2$.

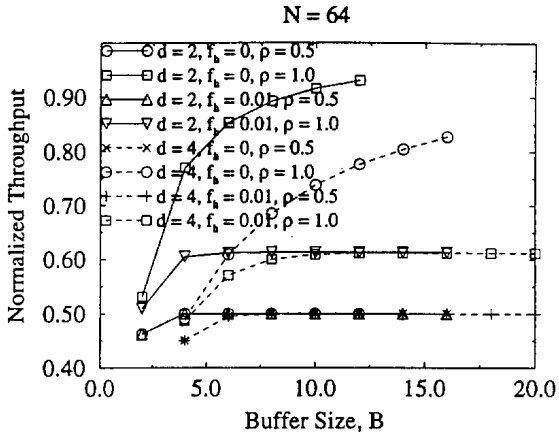

Figure 6: Normalized throughput versus B for $N = 64$.

performance of a network when d is large. The reason is that, when a tree saturation [21] takes place in a network, the cold outputs in the type 1 (hot type) SE are most affected by the phenomenon. The effect eases as switch type increases. For a particular network size N, the smaller the SE size (d), the greater the number of SE types in the last stage, the less is the effect of the hot spot traffic on the overall throughput of the network. Increasing the buffer size may alleviate the monopolization effect inside the network under low hot spot values, however, due to the properties of the shared buffer network, increasing the buffer size has very little effect on improving the performance of the network when hot spot traffic increases. For example, as shown in Fig. 6, buffer sizes greater than four have no effect on the throughput of the network for $d = 2$, $f_h = 0.01$, and $\rho = 1.0$. The impact of increasing the buffer size on the delay and packet loss under uniform and hot spot traffic is shown in Figs. 7 and 8.

A comparison between the proposed model and sim-

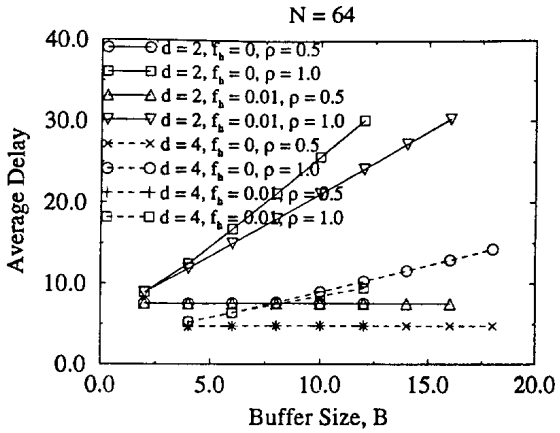

Figure 7: Average delay versus B for $N = 64$.

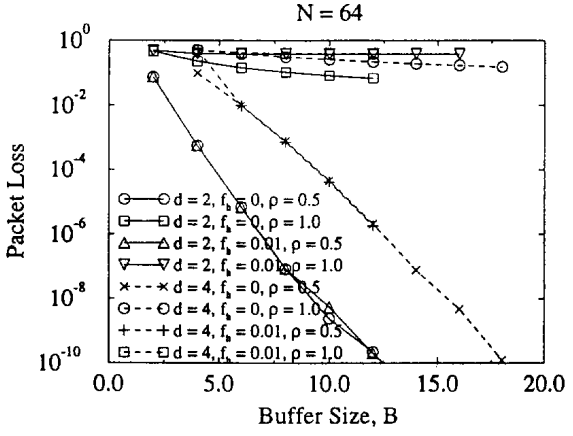

Figure 8: Packet loss versus B for $N = 64$.

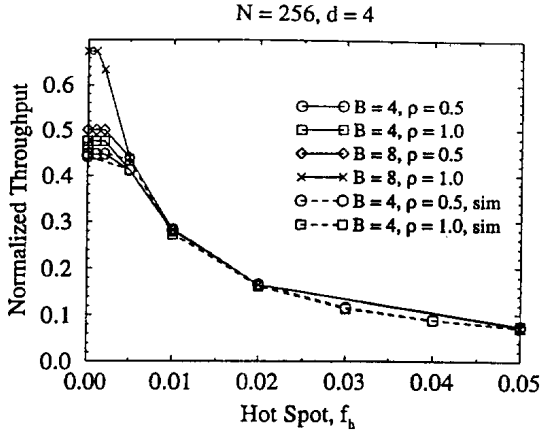

Figure 9: Normalized throughput versus f_h for $N = 256$.

ulation has been made in Fig. 9 for $N = 256$ and $d = 4$. The results obtained by the model are close to the simulation results under both uniform and hot spot traffic. The small inaccuracy of the model under uniform traffic is due to the fact that the model does not take the time correlation of the blocked packets [5] into account.

Figs. 10 and 11 illustrate hot, cold and total buffer occupancy, expressed as fractions of the buffer spaces occupied by the hot, cold and overall traffic in the first stage SE respectively. Under uniform traffic, buffer occupancy is proportional to the offered traffic of the network, all outlets taking a fair amount of the total buffer space. However, this proportion changes in favor of the hot traffic, when some hot spot value is introduced. When the hot spot is more than 0.1, the hot traffic saturates the buffers, even under an input load of as low as 0.4. Again, increasing the buffer size makes little improvement on this effect. We have reported similar results for different buffer size and hot spot values in [16]. Better buffer utilization is an advantage of a shared buffer network which improves the throughput as compared to networks using other buffer disciplines. However, under hot spot traffic, all of the buffers may be exhausted by the hot traffic. Fig. 12 contrasts a shared buffer an an output buffer network for $N = 64$, and $d = 2$. For a reasonable comparison, we have assumed the same number of buffer spaces (B) per SE for both architectures. The results for the output buffer network are obtained from a simulation program using a similar methodology to that described in Sec. 3. As shown in Fig. 12, a shared buffer network performs better under uniform traffic or when the hot spot value is small. Under high hot spot values, an output buffer network performs better. This can be explained as follows. In output buffering, the hot traffic degrades the throughput of the outputs which share the same buffers as a packet destined to them traverses through different stages. However, there are still some outputs that do not share any buffer with the hot output, and therefore are not affected by the hot traffic. Unlike output buffering, the hot spot traffic in a shared buffer affects all of

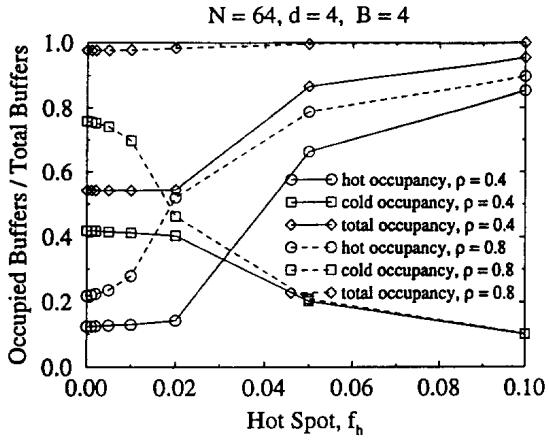

Figure 10: Ratio of hot, cold, and total buffer occupancy of the first stage SE for $N = 64$, $d = 4$, and $B = 4$.

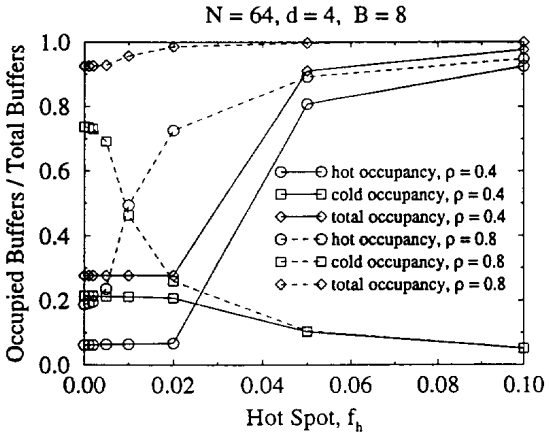

Figure 11: Ratio of hot, cold, and total buffer occupancy of the first stage SE for $N = 64$, $d = 4$, and $B = 8$.

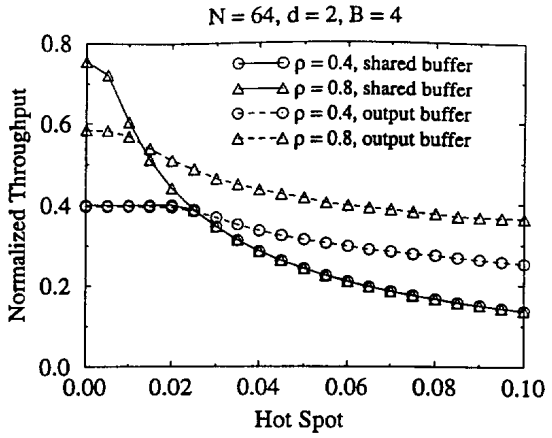

Figure 12: Comparison of the throughput of shared buffer and output buffer networks for $N = 64$, and $d = 2$.

the outputs, since all of them share at least the buffers at the first stage which, if congested, will degrade the throughput of all non-hot outputs as well.

Buffer monopolization in shared buffer networks can be minimized if a proper buffer management is utilized to limit the maximum number of buffers used by any outlet to some specified value.

5 Conclusion

We have developed an analytical model to study the performance of multistage networks constructed from shared buffer switching elements with an arbitrary SE size and buffer size. From the model, the throughput, packet delay, and packet loss probability in such networks have been derived, and various numerical results have been illustrated. We also have compared the results obtained from the model and computer simulations, and they have been found to be in close agreement. The model does not account for the correlation of packets in successive cycles. This allows a packet which is blocked during a cycle to bypass the congested route during subsequent cycles, giving rise to a higher throughput than simulation. In reality, a blocked packet in an SE always hunts for the same outlet of the SE during successive cycles. Under uniform traffic or under low hot spot values, a shared buffered network has better performance in terms of throughput, delay and packet loss as compared to a network with output buffering.

The proposed model can be used by network designers to study the effect of the different network parameters on the performance, and optimize the cost/performance ratio of the network. Furthermore, it can be easily modified to handle local flow control and other topologies of multistage networks.

References

[1] H. Yoon, K.Y. Lee, and M.T. Liu, "Performance analysis of mulitbuffered packet-switching networks in multiprocessor systems," *IEEE Trans-*

actions on Computers, vol. 39, no. 3, pp. 319–327, March 1990.

[2] S.H. Hsiao and R.Y. Chen, "Performance analysis of single-buffered multistage interconnection networks," *in Proc. Third IEEE Symposium on Parallel and Distributed Processing*, pp. 864–867, December 1991.

[3] M. Atiquzzaman and M.S. Akhtar, "Performance of buffered multistage interconnection networks in non uniform traffic environment," *in Proc. Seventh International Parallel Processing Symposium*, California, pp. 762–767, April 1993.

[4] M. Atiquzzaman and M.S. Akhtar, "Effect of non-uniform traffic on the performance of unbuffered multistage interconnection networks," *IEE Proceedings - Computer and Digital Techniques*, vol. 141, no. 3, pp. 169–176, May 1994.

[5] B. Zhou and M. Atiquzzaman, "Improved performance model of multibuffered multistage interconnection network under general traffic patterns," *IEEE INFOCOM '94: Conference on Computer Communications*, Toronto, Canada, pp. 1448–1455, June 1994.

[6] Y-C. Jenq, "Performance analysis of a packet switch based on single-buffered Banyan network," *IEEE Journal on Selected Areas in Communications*, vol. SAC-1, no. 6, pp. 1014–1021, December 1983.

[7] T. Szymanski and S. Shaikh, "Markov chain analysis of packet-switched Banyans with arbitrary switch sizes, queue sizes, link multiplicities and speedups," *in Proc. IEEE INFOCOM '89: 8th Annual Joint Conference of the IEEE Computer and Communication Societies*, Ontario, Canada, pp. 960–971, April 1989.

[8] Y.S. Yeh, M.G. Hluchyj, and A.S. Acampora, "The Knockout switch: A simple, modular architecture for high-performance packet switching," *IEEE Journal on Selected Areas in Communications*, vol. SAC-5, no. 8, pp. 1274–1283, October 1987.

[9] Y. Sakurai, N. Ido, S. Gohara, and N. Endo, "Large-scale ATM multistage switching network with shared buffer memory switch," *IEEE Communications Magazine*, pp. 90–96, January 1991.

[10] A. Monterosso and A. Pattavina, "Performance analysis of multistage interconnection networks with shared-buffered switching elements for ATM switching," *in Proc. IEEE INFOCOM '92: Conference on Computer Communications*, Florence, Italy, pp. 124–131, May 1992.

[11] V.S. Frost and B. Melamed, "Traffic modeling for telecommunications networks," *IEEE Communications Magazine*, pp. 70–81, March 1994.

[12] J.S. Turner, "Queueing analysis of buffered switching networks," *IEEE Transactions on Communications*, vol. 41, no. 2, pp. 412–420, February 1993.

[13] M. Saleh and M. Atiquzzaman, "Queueing analysis of shared buffer switches for ATM networks," *in Proc. GLOBECOM 94: IEEE Global Telecommunications Conference*, San Francisco, California, pp. 1070–1074, November 1994.

[14] H. Kuwahara, N. Endo, M. Ogino, T. Kozaki, Y. Sakurai, and S. Gohara, "A shared buffer memory switch for an ATM exchange," *in Proc. IEEE International Conference on Communications*, pp. 118–122, 1989.

[15] Y. Shobatake, M. Motoyama, E. Shobatake, T. Kamitake, S. Shimuzu, M. Noda, and K. Sakaue, "A one-chip scalable 8*8 ATM switch LSI employing shared buffer architecture," *IEEE Journal on Selected Areas in Communications*, vol. 9, no. 8, pp. 1248–1253, October 1991.

[16] M. Saleh and M. Atiquzzaman, "Buffer occupancy in ATM switches with single hot spot," *Electronics Letters*, vol. 31, no. 1, pp. 13–15, January 1995.

[17] J.H. Patel, "Processor-memory interconnections for multiprocessors," *in Proc. 6th Annu. Symp. on Comput. Arch.*, New York, April 1979.

[18] M. Saleh and M. Atiquzzaman, "Performance of shared buffer switches under non-uniform traffic pattern", *in Proc. Australian telecommunications and networking applications conference, ATNAC'94*, Melbourne, Australia, pp. 283-287, December 1994.

[19] G. Bianchi and J.S. Turner, "Improved queueing analysis of shared buffer switching networks," *IEEE/ACM Transactions on Networking*, vol. 1, no. 4, pp. 482–490, August 1993.

[20] S. Gianatti and A. Pattavina, "Performance analysis of shared-buffered Banyan networks under arbitrary traffic patterns," *in Proc. IEEE INFOCOM '93: Conference on Computer Communications*, pp. 943–952, March 1993.

[21] G.F. Pfister and V.A. Norton, "Hot spot contention and combining in multistage interconnection networks," *IEEE Transactions on Computers*, Vol. C-34, no. 10, pp. 943–948, October 1985.