Analysis of Window by-Pass Mechanism in Split Shared Buffer ATM Switches

Bin Zhou

Defence Commun. Research Centre
Institute for Telecommu. Research
University of South Australia
Adelaide 5095

M. Atiquzzaman

Dept. of Elect. & Comp. Systems Eng.
Monash University, Clayton
Melbourne 3168, Australia
atiq@eng.monash.edu.au

ABSTRACT

A Markov chain model for evaluating the performance of a multistage ATM switching fabric using split shared buffered switch elements (SE) is proposed. By comparing the results obtained from the proposed model and simulations, it is shown that the model is accurate in the presence of uniform and non-uniform traffic in the switch. To eliminate the influence of head of line blocking, a non-FIFO queueing discipline approach called a window by-pass mechanism is considered. By comparing the results obtained from the proposed model and previous models, the performance of ATM switches constructed of different buffered SEs is also studied.

1 INTRODUCTION

ATM (Asynchronous Transfer Mode) is the most promising technique to implement the transport network for Broadband ISDN (B-ISDN). Most of the ATM switches have buffers to resolve the blocking of ATM cells. There are various ATM switch architectures using the input, output, shared and crosspoint buffered SEs as shown in Fig 1. Many architectures have been studied [1, 2, 3] in an effort to realize a high-speed ATM switch with low cell loss probability. Each architecture has its own advantages and disadvantages and, therefore, it is very difficult to say which architecture is the best. Actually, the ATM switch fabrics already developed adopt various switch architectures.

Since cell loss is one of the critical issues in an ATM system, a switch architecture with low cell loss probability is required. A shared-buffer ATM switch shares the buffer among all output ports to achieve high buffer utilisation efficiency. However, the small access time required for the shared buffer becomes a bottleneck as the switch size increases. Moreover, shared buffer switches suffer from a phenomenon called *buffer hogging* in the case of a non-uniform traffic pattern. Therefore, the conventional approaches for the shared buffer switch realization would not be a general solution.

A split shared buffer architecture for high-speed ATM switch LSIs has been proposed in [4] in which multiple buffers are located between two crosspoint switches. By controlling the input-side crosspoint switch so as to equalise the utilisation rate of each buffer, these multiple buffers can be recognised as a single large shared buffers. Higher utilisation efficiency of buffer can thus be achieved, and the cell loss probability is minimised. By accessing the buffer memories in parallel via crosspoint switches, the time required to access the memories is greatly reduced. The *aim* of this paper is to develop a model which can be applied to evaluate the performance of ATM switch using split shared buffers under uniform and hotspot traffic patterns. We also compare the performance of ATM switches with output, crosspoint and split shared buffering schemes.

In Section 2, we briefly review the window by-pass mechanism in a buffered switch. In Section 3, we present a non HOL blocking model. Section 4 presents a variety of performance curves. Finally, in Section 5, we conclude the paper.

2 WINDOW BY-PASS MECHANISM

Head-of-line blocking can arise if the split shared buffered switch is operated under the first-in-first-out (FIFO) rule. To avoid this blocking, we employ a window by-pass mechanism [5] which however relaxes the FIFO rule. Computer simulations are performed to choose the window parameter for the model. Fig. 2 shows the plots of the throughput versus the traffic load with parameters of w = 1,2,3,4,5,6, and 7 where w indicates the

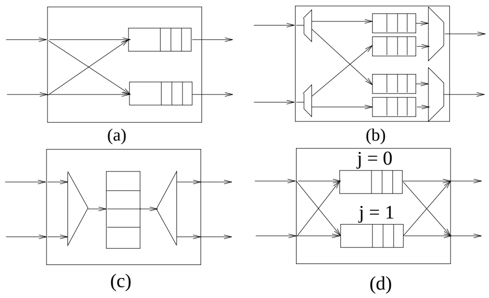

Figure 1: Buffering architecture (a) output (b) cross-point (c) usual shared (d) split shared

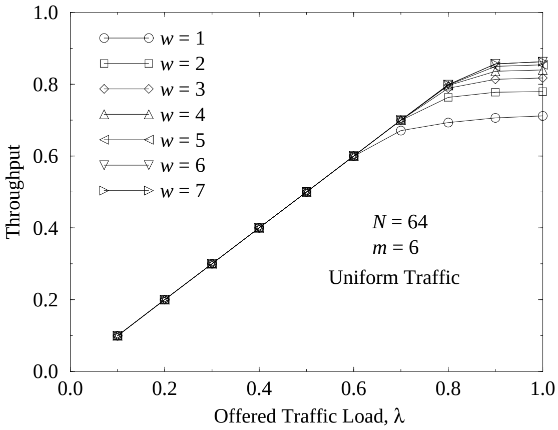

Figure 2: Throughput versus load obtained from simulation.

window size. Increasing w gives remarkable improvement of throughput and mean delay. Fig. 2 indicates that the throughput with $w = 3$ almost reaches its ideal value ($w = 6$). Note that in the case of $w = 6$, no blocking occurs, and the throughput is only limited by the buffer size and traffic pattern of the input ATM cells.

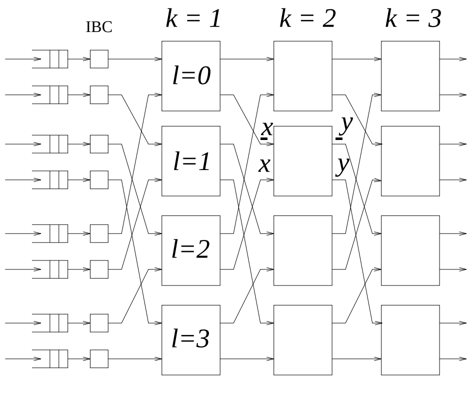

Figure 3: A three stages ATM switch fabric.

3 THE MODEL

We make the same assumptions as in [2] regarding the operation and environment of the switch. e.g., $N = 2^s$ inputs and N outputs; *synchronously* operation; *backpressure* mechanism; *Bernoulli* cell arrival process and uniform and hot spot traffic pattern. A non HOL blocking analytical model to evaluate the performance of an ATM switch using split shared buffered SEs with window size of three or more is developed.

Figure 3 shows a three stages ATM switch fabric and how the SEs are specified by parameters k, l, input ports x and $\underline{x}$, and output ports y and $\underline{y}$. Figure 4 shows three consecutive SEs in successive stages of the switch. We denote the input and output ports of the l-th SE at stage k by klx, $kl\underline{x}$, kly and $kl\underline{y}$. Let $(k+1)\hat{l}\hat{y}$ denote the input port of the $\hat{l}$-th SE, at the $(k+1)$-th stage, which is fed by the output port kly at the k-th stage. Also, let $(k-1)\tilde{l}\tilde{x}$ denote the output port of the $\tilde{l}$-th SE, at the $(k$ -1)-th stage, which feeds the input port klx at the k-th stage. j, $j \in \{0, 1\}$, denotes the buffer position. $j = 0$ means the upper buffer and $j = 1$ means the lower buffer. The following variables to be used in the model are defined.

$B^i_{klj}(t)$ = the probability that there are i cells in the buffer at klj at the beginning of cycle t, $j \in \{0, 1\}$

π_{ij} = the transition probability of a buffer from state i to state j, given that it is in state i.

m = the number of buffers in each SE.

s = the number of switching stages in the switch.

w = the window size.

$P_{kljv}(t)$ = the *static routing probability* that a cell, in the buffer klj, $j \in \{0, 1\}$, during clock cycle t, is destined to klv, $v \in \{y, \underline{y}\}$.

$Q_{klu}(t)$ = the probability that a cell is ready to come to port klu, $u \in \{x, \underline{x}\}$, during clock cycle t.

$C^i_{klj}(t)$ = the probability that i, $0 \le i \le 2$, cells are ready to come to the buffer at klj, $j \in \{0, 1\}$, during clock cycle t.

$a^i_{klj}(t)$ = the probability that i, $0 \le i \le 2$, cells in buffer klj advances to the next stage during clock cycle t, given that there are i cells in the buffer at klj, $j \in \{0, 1\}$.

$r_{kluj}(t)$ = the probability that a cell from port klu, $u \in \{x, \underline{x}\}$, advances to the buffer klj during

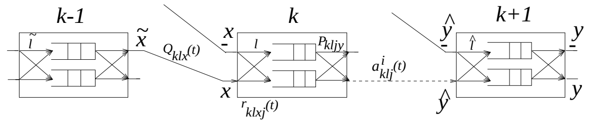

Figure 4: Illustration of $r_{klxj}(t)$, and $a^i_{klj}(t)$.

clock cycle t, given that a cell, destined for klj, $j \in \{0,1\}$, is ready to come to klu during the clock cycle t.

λ = the probability of a cell arriving at an IBC during a cycle.

$\overline{z} = 1 - z$ is defined for any variable z, and the steady state value of any variable $z(t)$ will be represented by z.

3.1 Routing Probabilities

The relationship between P_{kljy}, r_{klxj}, a^i_{klj} and Q_{klx} is shown in Figure 4. The probability that a cell, ready to come to the klu port and destined to klj buffer, can pass from klu to klj buffer is given by

$$\begin{gathered} r_{kluj} = 1 - \{B^m_{klj} + B^{m-1}_{klj}\} + B^m_{klj} a^2_{klj} \\ + B^{m-1}_{klj} a^1_{klj} + \{1 - Q_{kl\underline{u}} + 0.5 Q_{kl\underline{u}} + 0.25 Q_{kl\underline{u}}\} \\ \{B^{m-1}_{klj} a^0_{klj} + B^m_{klj} a^1_{klj}\},\ 1 \le k \le s \end{gathered}$$

The probability of departure of i cells from a buffer in a stage is determined by the possibility of conflict with a cell at the same stage and the availability of buffer space at the destined SE of the next stage. a^i_{klj}, $1 \le k \le s$, can then be obtained from r_{kluj} as follows:

$$\begin{gathered} a^1_{klj} = P_{kljy}\{0.5 r_{(k+1)\hat{l}\hat{y}0} + 0.5 r_{(k+1)\hat{l}\hat{y}1}\}\{(0.5 P_{kl\bar{j}y} \\ + P_{kl\bar{j}\underline{y}}) \textstyle\sum_{i=1}^m B^i_{kl\bar{j}}\} + P_{klj\underline{y}}\{0.5 r_{(k+1)\hat{l}\underline{\hat{y}}0} \\ + 0.5 r_{(k+1)\hat{l}\underline{\hat{y}}1}\}\{(0.5 P_{kl\bar{j}\underline{y}} + P_{kl\bar{j}y}) \textstyle\sum_{i=1}^m B^i_{kl\bar{j}}\} \\ a^2_{klj} = \{B^0_{kl\bar{j}} + 0.5 P_{klj\underline{y}} P_{kl\bar{j}\underline{y}} \textstyle\sum_{i=1}^m B^i_{kl\bar{j}} + 0.5 P_{kljy} \\ P_{kl\bar{j}y} \textstyle\sum_{i=1}^m B^i_{kl\bar{j}}\}\{0.5 r_{(k+1)\hat{l}\hat{y}0} + 0.5 r_{(k+1)\hat{l}\hat{y}1}\} \\ \{0.5 r_{(k+1)\hat{l}\underline{\hat{y}}0} + 0.5 r_{(k+1)\hat{l}\underline{\hat{y}}1}\} \\ a^0_{klj} = 1 - a^2_{klj} - a^1_{klj} \end{gathered}$$

$Q_{klu}(t)$ can be expressed as

$$\begin{aligned} Q_{klu}(t) &= \{1 - B^0_{(k-1)\tilde{l}0}(t-1)\} P_{(k-1)\tilde{l}0\tilde{u}} \\ &\quad + \{1 - B^0_{(k-1)\tilde{l}1}(t-1)\} P_{(k-1)\tilde{l}1\tilde{u}} \\ Q_{1lu}(t) &= 1 - B^0_{0\tilde{l}1}(t-1) \end{aligned}$$

where $B^0_{0\tilde{l}1}(t-1)$ is the probability of the IBC being empty at $t-1$. The probabilities of cells coming

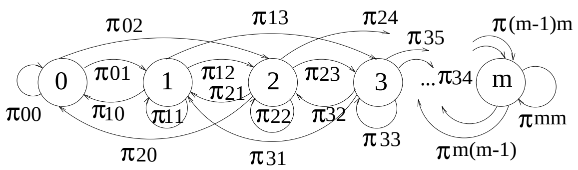

Figure 5: Markov state transition for stage buffers.

to the klj buffer are given by

$$\begin{aligned} C^0_{klj} &= \{(1 - Q_{kl\underline{u}}) + 0.5 Q_{kl\underline{u}}\} \\ &\quad \{(1 - Q_{klu}) + 0.5 Q_{klu}\} \\ C^1_{klj} &= 0.5 Q_{klu}\{1 - Q_{kl\underline{u}} + 0.5 Q_{kl\underline{u}}\} \\ &\quad + 0.5 Q_{kl\underline{u}}\{1 - Q_{klu} + 0.5 Q_{klu}\} \\ C^2_{klj} &= 0.25 Q_{klu} Q_{kl\underline{u}} \end{aligned}$$

3.2 Stage Buffers

A buffer at any SE at stage k, $1 \le k \le s$, is modelled as a Markov chain. The state of a buffer is represented by the number of cells in the buffer. The $(m+1)$ state Markov chain of the buffer is shown in Figure 5, with $\pi_{ij}, i, j = 0, 1, \ldots, m$ being the transition probability from state i to state j. With these considerations, the transition probabilities, $\mathbf{\Pi}_{klj} = [\pi_{i,j}, 0 \le i, j \le m]$ at buffer klj are described by the following transition equations.

$$\mathbf{\Pi}_{klj} = \begin{bmatrix} \pi_{00} & \pi_{01} & \cdots & \pi_{0m} \\ \pi_{10} & \pi_{11} & \cdots & \pi_{1m} \\ \vdots & \vdots & \cdots & \vdots \\ \pi_{m0} & \pi_{m1} & \cdots & \pi_{mm} \end{bmatrix}$$

where, for $1 \le i \le m-1$ and $i + 3 \le j \le (i-2) \& (i \ge 2)$

$$\begin{aligned} \pi_{00} &= C^0_{klj} \\ \pi_{01} &= C^1_{klj} \\ \pi_{02} &= C^2_{klj} \\ \pi_{03} &= 0 \\ &\vdots \\ \pi_{10} &= C^0_{klj} a^1_{klj} \\ \pi_{11} &= C^1_{klj} a^1_{klj} + C^0_{klj} a^0_{klj} \\ \pi_{12} &= C^1_{klj} a^0_{klj} + C^2_{klj} a^1_{klj} \\ \pi_{13} &= C^2_{klj} a^0_{klj} \\ \pi_{14} &= 0 \\ &\vdots \\ \pi_{i(i-2)} &= C^0_{klj} a^2_{klj} \\ \pi_{i(i-1)} &= C^0_{klj} a^1_{klj} + C^1_{klj} a^2_{klj} \\ \pi_{ii} &= C^0_{klj} a^0_{klj} + C^1_{klj} a^1_{klj} + C^2_{klj} a^2_{klj} \end{aligned}$$

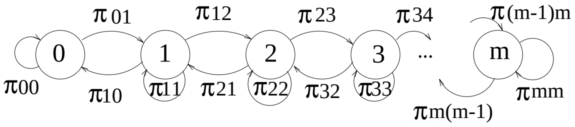

Figure 6: Markov state transition for IBC buffers.

$$
\begin{aligned}
\pi_{i(i+1)} &= C^1_{klj}a^0_{klj} + C^2_{klj}a^2_{klj} \\
\pi_{i(i+2)} &= C^2_{klj}a^0_{klj} \\
\pi_{ij} &= 0 \\
&\vdots \\
\pi_{mm} &= \overline{r}_{klj} + a^1_{klj}(C^1_{klj} + C^2_{klj})
\end{aligned}
$$

for $j \le m-2$

$$\pi_{mj} = 0$$

The steady buffer state probabilities, $\mathbf{B}_{klj} = \left[B^0_{klj}, B^1_{klj}, \ldots, B^m_{klj}\right]$, for buffer klj are obtained by solving

$$\mathbf{B}_{klj} = \mathbf{B}_{klj}\mathbf{\Pi}_{klj}$$

with boundary conditions, $\sum_{i=0}^{m} B^i_{klj}(t) = 1$.

For example, the balance equation for $B^i_{klj}(t)$ can be obtained from the above equation as follows:

$$
\begin{gathered}
B^i_{klj}(t) = C^2_{klj}(t)B^{i-2}_{klj}(t)a^0_{klj}(t) + B^{i-1}_{klj}(t) \\
\{C^1_{klj}(t)a^0_{klj}(t) + C^2_{klj}(t)a^1_{klj}(t)\} + B^i_{klj}(t) \\
\{C^2_{klj}(t)a^2_{klj}(t) + C^1_{klj}(t)a^1_{klj}(t) + C^0_{klj}(t)a^0_{klj}(t)\} \\
+B^{i+1}_{klj}(t)\{C^0_{klj}(t)a^1_{klj}(t) + C^1_{klj}(t)a^2_{klj}(t)\} \\
+B^{i+2}_{klj}(t)C^0_{klj}(t)a^2_{klj}(t), \;\; 2 \le i \le m-1, m \ge 5
\end{gathered}
$$

3.3 IBC Buffers

Each IBC buffer is of size m and is modelled as a Markov chain. The probability of departure (a^i_{0lj}) of a cell is determined by the availability of buffer space at the first stage, and is given by

$$
\begin{aligned}
a^1_{0lj} &= \{B^0_{0l\overline{j}} + (0.5P_{kl\overline{j}y} + P_{kl\overline{j}\underline{y}})\sum_{i=1}^{m} B^i_{kl\overline{j}}\} \\
&\quad \{0.5r_{1\hat{l}\hat{j}j} + 0.5r_{1\hat{l}\hat{j}\overline{j}}\} \\
a^0_{0lj} &= 1 - a^1_{0lj}
\end{aligned}
$$

The $m+1$ state Markov chain model of an m-buffer IBC buffer is shown in Figure 6. The transition probabilities, $\mathbf{\Pi}_{0lj} = [\pi_{i,j}, 0 \le i, j \le m]$, are given by

$$
\begin{bmatrix}
\overline{\lambda} & \lambda & 0 & \ldots & 0 \\
\overline{\lambda}a^1_{0lj} & \lambda a^1_{0lj} + \overline{\lambda}a^0_{0lj} & \lambda a^0_{0lj} & \ldots & 0 \\
0 & \overline{\lambda}a^1_{0lj} & \overline{\lambda}a^0_{0lj} + \lambda a^1_{0lj} & \ldots & 0 \\
0 & 0 & \overline{\lambda}a^1_{0lj} & \ldots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & 0 & \ldots & \lambda a^0_{0lj} \\
0 & 0 & 0 & \ldots & 1
\end{bmatrix}
$$

The steady buffer state probabilities of the IBC are obtained by solving the Markov chain. The steady IBC state probabilities, $\mathbf{B}_{0lj} = \left[B^0_{0lj}, B^1_{0lj}, \ldots, B^m_{0lj}\right]$, are obtained by $\mathbf{B}_{0lj} = \mathbf{B}_{0lj}\mathbf{\Pi}_{0lj}$

The objective of the analysis is to determine the values of B^i_{klj}, the steady state probability of buffer occupancy at the k-th stage and a^i_{klj}, the probability of departure. Since the equations describing the dynamics of the switch are described by recurrence relations, the solution is obtained by an iterative method [2].

3.4 MEASURES

By putting together all of the above parameters, throughput, cell loss and delay can be computed. The throughput (μ_{slj}) of the switch at the output slv is given by

$$\mu_{slj} = 1 - \lim_{t\to\infty} B^0_{slj}(0,t)$$

The normalised throughput (μ) is given by the number of cells being output from a switch output link during a cycle, and can be given by

$$\mu = \frac{\sum_{\text{all output}} \mu_{slj}}{N}$$

The probability of a cell being lost at the switch input is given by $\eta_{0lj} = B^m_{0lj}$. The probability of a cell loss for uniform traffic can be given by:

$$\eta = \frac{\lambda - \mu}{\lambda}$$

Cell delay is defined to be the number of clock cycles a cell takes to reach the output port from a switch input. Let δ_{klj} be the cell delay in the buffer of an SE in stage k.

$$\delta_{klj} = \lim_{t\to\infty} \frac{\sum_{i=1}^{m} iB^i_{klj}(t)}{\sum_{i=1}^{m} B^i_{klj}(t)\{a^1_{klj}(t) + a^2_{klj}(t)\}}$$

The cell delay in the IBC is given by

$$\delta_{IBC} = \lim_{t\to\infty} \frac{\sum_{i=1}^{m} iB^i_{0lj}(t)}{\sum_{i=1}^{m} B^i_{0lj}(t)a^1_{0lj}(t)}$$

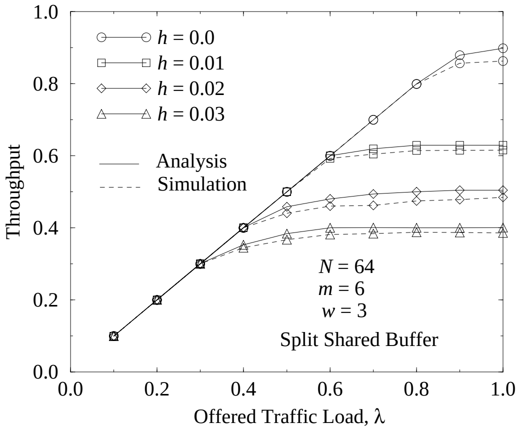

Figure 7: Throughput versus offered load, N=64.

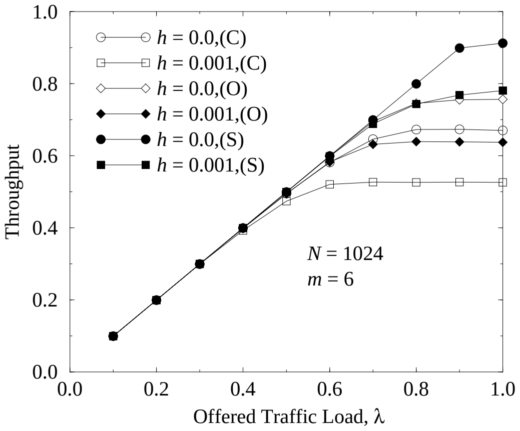

Figure 9: Comparison of the throughput.

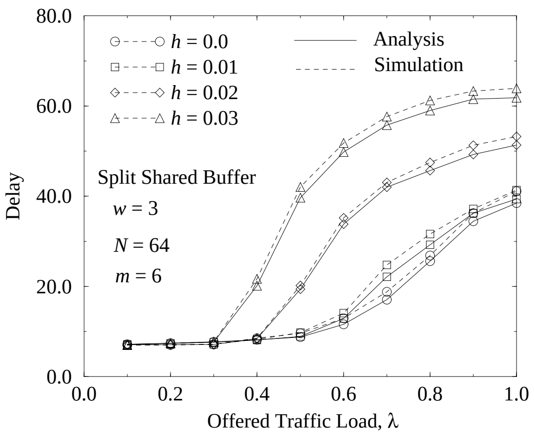

Figure 8: Delay versus offered load, N=64.

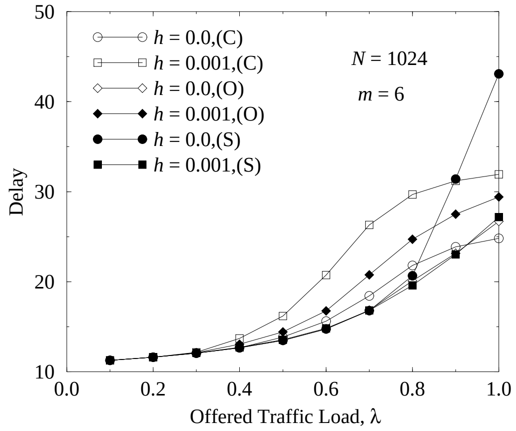

Figure 10: Comparison of the delay.

The cell total delay in the switch including the IBC buffer is therefore given by

$$\delta \quad = \quad \delta_{IBC} + \sum_{k=1}^{s} \delta_{klj}$$

4 RESULTS

In this section, the results from the proposed model will be compared with those obtained from simulation, output buffer model [2] and crosspoint buffer model [3]. In the figures, (O), (S) and (C) stand for the Output buffering, Split shared buffering and Crosspoint buffering, respectively.

In Figs. 7 and 8, the results from the model and simulation are compared. The throughput and delay of the network are plotted versus offered load (λ) where the probabilities of the hot spot (h) takes values of 0.0 (uniform traffic), 0.01, 0.02 and 0.03. In these examples, buffer size is assumed to be 6. The switch size (N) is 64. While the throughput is improved when h decreases, and the network delay rises when h increases. The curves for simulation and the analytical model are very close. Fig. 7 shows that when the input load is less than the maximum throughput, the throughput is proportional to input load. When the input load approaches the maximum throughput, increase of input load does not significantly affect the switch throughput. However, the transmission delay will increases as input load increases (see Fig. 8). In Figs. 9 and 10, the throughput and delay of the switch for three buffering schemes are compared for hot spot probabilities of 0.0 (uniform traffic) and 0.001. In these examples, the buffer size is assumed to be 6. The switch size (N) is 1024. Split shared buffering has a lower average delay. In addition, split shared buffering has a higher throughput when the offered traffic load is high. The output buffer performs better than the crosspoint buffer. Under uniform traffic (see Fig. 11), the load is uniformly distributed among the buffers in the same stage. The buffer occupancy of the buffers in the same stage is identical, and buffer occupancy decreases with every stage of the switch. It is shown that the split shared buffer switch shares the buffer among all output ports to achieve high buffer utilisation. Fig. 12 shows the throughput with respect to the number of buffers in the uniform traffic sit-

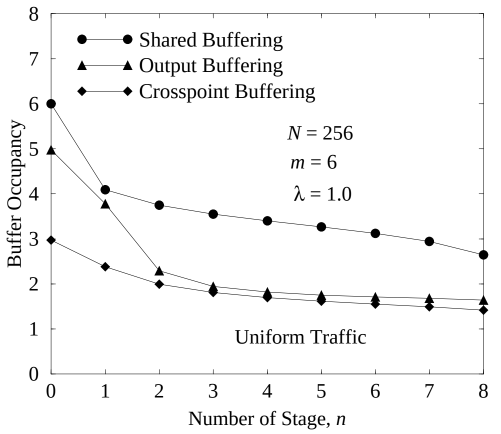

Figure 11: Comparison of the buffer occupancy.

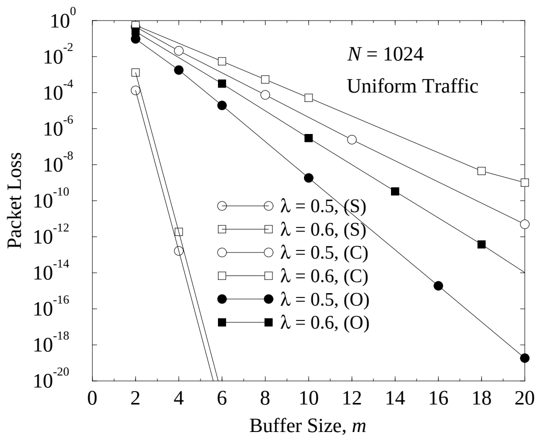

Figure 13: Cell loss versus buffer size.

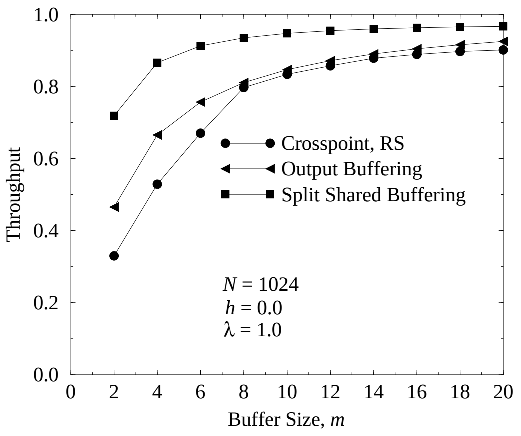

Figure 12: Throughput versus buffer size.

uations. The offered traffic load (λ) is kept at full load. The switch size (N) is 1024. The split shared buffer gives higher throughput than crosspoint and output buffers. Note that the throughput saturates and does not increase when the buffer size is more than 10. Fig. 13 shows the cell loss probability at the switch against the number of buffers for three buffering schemes, where the switch size (N) is 1024 and the input load (λ) is varied from 0.5 to 0.6. In the split shared buffer switch, low cell loss can be achieved because the probability of a blocked cell is very small. To achieve a packet loss probability of 10^{-8} with an offered load of 0,6, a buffer size of four is required for the split shared buffering, 12 for output buffering and 18 for cross-point buffering. The cell loss probability decreases with increasing number of buffers. It decreases slowly at a heavy offered traffic load but decreases sharply at a small offered traffic load. As can be seen, the cell loss probability is very low for the split shared buffering when compared with those of output and crosspoint buffering schemes.

5 CONCLUSIONS

An analytical method for a split shared buffered ATM switch is presented. A careful evaluation of the analytical results is made by comparing them with simulation results. Performance of the split shared buffer switch is studied by comparing the results from output, cross-point and split shared buffering. The cell loss probability of the split shared buffering scheme is much lower than that of the output-buffering or cross-point buffering scheme for the same amount of buffer in a SE.

6 ACKNOWLEDGMENTS

The work reported in this paper was supported by Telstra, Siemens, Monash University, RDN-CRC and DSTO.

References

[1] B. Zhou and M. Atiquzzaman, "A performance comparison of buffering schemes for multistage switches," *ICA3PP'95, First International Conference on Algorithms And Architectures for Parallel Processing*, Brisbane, Australia, pp. 809–818, April 19-21, 1995.

[2] B. Zhou and M. Atiquzzaman, "Performance analysis of finite output-buffered multistage ATM switching fabrics," *IEEE INFOCOM'95*, Massachusetts, U.S.A., pp. 16–23, April 2-6, 1995.

[3] B. Zhou and M. Atiquzzaman, "Performance ATM switch fabric using cross-point buffers," *Asia-Pacific Conference on Communications*, Osaka, Japan, June 14-16, 1995.

[4] H. Kondoh, H. Notani, H. Yamanaka, K. Higashitani, and H. Saito, "A shared multibuffer

architecture for high-speed ATM switch LSIs," *IEICE Trans. Electron.*, vol. E76-C, no. 7, pp. 1094–1101, July 1993.

[5] M.G. Hluchyj and M.j. Karol, "Queueing in high-performance packet switching," *IEEE Journal on Selected Areas in Communications*, vol. 6, no. 9, pp. 1587–1597, December 1988.