A Parallel-Tree Switch Architecture for ATM Networks

Mayez Al-Mouhamed, Habib Youssef, and Wasif Hasan

Computer Engineering Department
College of Computer Science and Engineering
King Fahd University of Petroleum and Minerals
Dhahran 31261, Saudi Arabia.
mayez,habib,wasif@ccse.kfupm.edu.sa

Abstract

In this paper we present a novel ATM switch called *Parallel-Tree Banyan Switch Fabric* (PTBSF) that consists of parallel Banyans arranged in a tree topology. Packets enter at the topmost Banyan. Internal conflicts are eliminated by using a conflict-free 3×4 switching element which distributes conflicting cells over different Banyans. Thus, cell loss may occur only at the lowest Banyan. Increasing the number of Banyans leads to noticeable decrease in the cell loss rate. The switch can be engineered to provide arbitrarily high throughput and low cell loss rate without the use of input buffering nor cell pre-processing. The performance of the switch is evaluated analytically under uniform traffic load and by simulation under a variety of ATM traffic loads. Compared to other proposed architectures, the switch exhibited stable and excellent performance with respect to cell loss and switching delay for all studied conditions as required by ATM traffic sources. The advantages of PTBSF are modularity, regularity, self-routing, low processing overhead, high throughput and robustness under a variety of ATM traffic conditions.

1 Introduction

Modern telecommunication networks are evolving at a fast pace. Many newly emerging applications have diverse service features and require huge bandwidth. The Asynchronous Transfer Mode (ATM) is the transfer mode of B-ISDN. ATM technology is rapidly gaining ground. ATM values itself as a technology capable of scaling up smoothly and efficiently to increased demands for higher transmission speeds and larger transfer loads. An ATM switch must be capable of supporting diverse traffic types and accommodating bit rates on the order of 100 Mbps and higher per input port. Thus, tremendous efforts are being made in order to design and implement a switch which is capable of switching cells at an extremely high rate and can handle diverse traffic types with least delay and extremely small cell loss rate [1].

The use of internal buffering in the Buffered-Banyan [2] increases the switch complexity and head of line (HOL) queuing delays. Internal blocking can be avoided if we use a sorting network in front of the routing network but the complexity of the sorter becomes the problem[3]. An alternative strategy [4] to increase throughput is to distribute the incoming traffic over parallel Banyans, and successfully routed packets in distinct Banyans are forwarded to the corresponding output buffers. In this case, the throughput reaches some saturation where increase in the number of planes no longer produces increase in throughput. The throughput of parallel Banyans [5] can be improved when separate control plane and data planes are used together with input buffering for conflicting cells.

The tandem Banyan switching fabric (TBSF) [6] consists of arranging Banyans in series. When a conflict occurs, one of the packets is routed correctly while the other is misrouted. At the output of each Banyan, those packets which have been routed properly are forwarded to the output buffers while the misrouted packets go into the next Banyan. The main drawback of TBSF is the large delay jitter. The piled Banyan switching fabric (PBSF) [7] uses multiple Banyan networks arranged one above the other, and constructed with 4×4 switches. If two packets conflict in some Banyan, one packet is correctly routed and the other is routed to a switching element in the next lower Banyan plan. At any level, a loss can occur if three packets compete for the same output. The throughput of the piled Banyan saturates at 98% under full load regardless of the number of Banyans.

In this paper, we present a Banyan switch called the *Parallel-Tree Banyan Switching Fabric* (PTBSF). PTBSF has simple control, self-routing, no internal buffering, and no recirculation to minimize switching delay and reduce hardware as possible. Its performance scales up well with the addition of reasonable extra hardware. The cell loss rate must meet ATM conditions. The switch has a regular architecture to facilitate VLSI implementation.

The paper is organized as follows. In Section 2 we describe the proposed switch architecture, operations, hardware complexity, and switching delay. In section 3, we discuss the switch performance under uniform traffic conditions. Then in section 4, we study the switch performance under a simulated ATM workload. We conclude in section

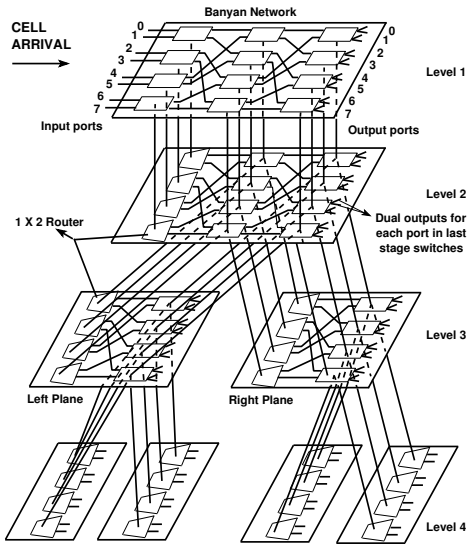

Figure 1: The parallel-tree switch architecture

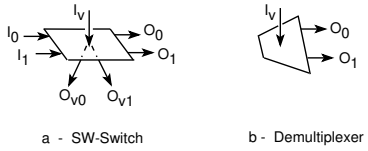

Figure 2: Switching elements

5.

2 The parallel-tree Banyan switch fabric

The PTBSF consists of Banyan networks arranged in a *parallel-tree structure* from layer 2 onwards as shown in Figure 1. Each Banyan network within the switch consists of 3×4 switching elements (SW) as shown in Figure 2-a. Each SW has two horizontal inputs (I_0 and I_1) and one vertical input (I_v). There are 2 horizontal outputs (O_0 and O_1) and 2 vertical outputs (O_{v0} and O_{v1}).

The packets arrive at the inputs of the topmost Banyan. When a conflict occurs between two packets in the first level, one of the packets is routed correctly, while the other is routed vertically downward to the corresponding switching element in the next lower level. If two cells arrive at a switching element, destined to the same output, one cell is routed to the correct output while the other is routed through O_{v0} to a switching element in the same position in the next layer Banyan. From level two onwards, there can be three cells at a switch input requesting the same output. In this case, the switch routes one cell to O_0 or O_1 and the other two cells are routed vertically through the outputs O_{v0} and O_{v1} to the corresponding switching elements in the next lower layer Banyans. Cell loss occurs only at the vertical outputs of the lowest layer. Forwarding the cell vertically downward instead of misrouting it preserves the routing achieved up to that stage.

From the second layer onwards, there can be no cells forwarded vertically downward at the first stage. Therefore, from layer two onwards, each Banyan network is made up of two types of switching elements. The first stage consists of 1×2 *Demultiplexer* (DM) (see Figure 2-b), while the other stages consist of 3×4 switching elements (SW).

2.1 Hardware resource requirement

A PTBSF with L layers has 2^{L-1} Banyan planes. However, the basic element of the first stage for a Banyan at layer $l \geq 2$ is a 1×2 demultiplexer DM (see Figure 2-b). In general, each Banyan of Layer l, $l \geq 2$, has one stage that consists of simple DM switches, and $n - l + 1$ stages made of SW's. A DM has much less hardware than an SW. Therefore, we shall be using the number of SW's as a hardware complexity metric of the PTBSF switch. It can be proved that a PTBSF with $N = 2^n$ inputs and L layers requires $S_{PTBSF} = (n - L + 2)2^{L-1} - 1$ SW-stages and $2^{L-1} - 1$ DM-stages. A TBSF with 2^n inputs and L Banyans in tandem requires $S_{TBSF} = nL$ *SW-stages*. For a PBSF with 2^n inputs and L layers, the number of SW-stages in level l is equal to $n-l+1$. Therefore, the overall number of SW-stages in all L levels is equal to $S_{PBSF} = \Sigma_{l=1}^{L}(n - l + 1)$, which simplifies to $S_{PBSF} = L(2n - L + 1)/2$.

Since the SW-switch dominates the hardware, it is important to evaluate the effective number of $n-$stage Banyans constructed with SW-switches only. The effective number of SW-Banyans in each of the TBSF, PBSF, and PTBSF is $E_{TBSF}(n, L) = L$, $E_{PBSF}(n, L) = (2n - L + 1)L/2n$, and $E_{PTBSF}(n, L) = ((n - L + 2)2^{L-1} - 1)/n$, respectively.

The total number of vertical interconnections of the PTBSF is $W^v_{PTBSF} = 2^{n-1}[2^{L-1}(n - L + 3) - n - 2]$. The total number of horizontal interconnections is $W^h_{PTBSF} = 2^n[2^{L-1}(n-L+4)-2]$. Therefore, the overall number of required interconnections is $W_{PTBSF} = W^h_{PTBSF} + W^v_{PTBSF}$ which simplifies to $W_{PTBSF} = 2^{n-1}[2^{L-1}(3n - 3L + 11) - n - 6]$.

For TBSF and PBSF, the overall number of required interconnections are $W_{TBSF} = (n + 1)L2^n$ and $W_{PBSF} = 2^n[(L - 2)(2n - L + 1) + 3n]$, respectively. For example, L must be 5 to achieve a cell loss of 10^{-6} with $N = 64$ for the TBSF and PTBSF requires 6.2 effective SW-Banyans which correspond to $L = 5$. In this case, the number of interconnections required by TBSF is 2240 and the number required by PTBSF is 6784, which is about 3 times more than TBSF. In general, PTBSF always requires 2 to 3 times the number of interconnections required by TBSF.

In the following, we analyze the switching delay in each of PTBSF, PBSF, and TBSF. The state of the basic switch in PTBSF is a boolean function of the form $ST_{PTBSF} = \Sigma d_2 d_1 d_0 pr_2 pr_1 pr_0$, where d_2, d_1, and d_0 are the destination bits of the incoming cells and pr_2, pr_1, and pr_0 are the cell priorities. Using dual input gates, the number of gate levels (n_{gate}) required to implement a sum of product $\Sigma b_1.b_2.\ldots.b_v$ is $n_{gate} = log_2(2^{n_v} \times n_v)$, where n_v is the number of boolean variables per *minterm*. Evaluating the state of a switch in the PTBSF requires $log_2(2^6 \times 6)$ gate delays

Switch	Cells	$P(O_0)$ or $P(O_1)$	$P(O_{vo})$	$P(O_{v1})$	P(loss)
PTBSF	One cell	$(1-p)(1-q)p+(1-p)^2q/2$	0	0	0
	Two cells	$3(1-q)p^2/4+3(1-p)pq/4$	$(1-q)p^2/2+(1-p)pq$	0	0
	Three cells	$7p^2q/8$	p^2q	$p^2q/4$	0
TBSF	One cell	$p(1-p)$	-	-	0
	Two cells	$3p^2/4$	-	-	$p^2/2$
PBSF	One cell	$(1-p)(1-q)p+(1-p)^2q$	0	0	0
	Two cells	$3(1-q)p^2/4+2(1-p)pq$	$(1-q)p^2/4+(1-p)pq$	$P(O_{vo})$	0
	Three cells	p^2q	$3p^2q/4$	$3p^2q/4$	$p^2q/4$

Figure 3: Analytical models under uniform traffic

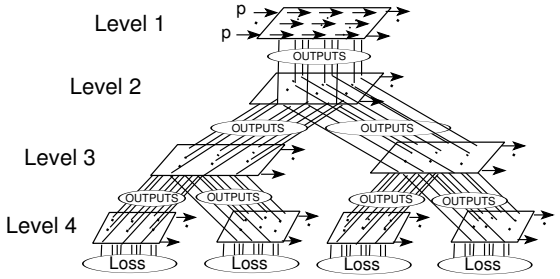

Figure 4: Computation of the analytical model for PTBSF

which gives 8 gate levels. Also, one latch is needed to store the state. The state of the 4×4 switch of PBSF is of the form $ST_{PBSF} = \Sigma d_2 d_1 d_0 p r_2 p r_1 p r_0$ for each pair of horizontal and vertical outputs, which indicates that 8 gate levels and one latch are needed for each pair. The longest path from switch input to output buffers in each of PTBSF and PBSF consists of $n+L-1$ switching elements when there are $N=2^n$ inputs and L levels. Therefore, the approximate maximum switching delay of PTBSF and PBSF is $T_{PTBSF} = T_{PBSF} = (n+L-1)(8T_{gate}+T_{latch})$, where T_{gate} and T_{latch} are the gate and latch delays, respectively. Similarly, the delay in switching TBSF is $T_{TBSF} = nL(6T_{gate}+T_{latch})$. Therefore, the switching delay of PTBSF (as well as that of PBSF) grows linearly with $n+L$, whereas the delay of TBSF grows with the product $n \times L$.

For example, to achieve a cell loss of 10^{-6} under uniform traffic with $N=64$, we must have $L=10$ for TBSF [6] and $L=5$ (or 6.2 effective SW banyans) for PTBSF (Figure 5). Under these conditions, $T_{TBSF}=360$ and $T_{PTBSF}=80$ time units if one excludes the latch delay. If we increase N to 256, then $T_{TBSF}=576$ and $T_{PTBSF}=104$ time units. Hence, the switching delay of PTBSF is much smaller due to its parallel Banyan structure.

3 Throughput performance under uniform traffic

We assume a time slotted synchronized operation of the switch, where the slot size is greater than or equal to the switch processing time. For each input and each time slot, a cell is issued with probability p and cells destinations are uniformly distributed over all outputs.

We derive analytical expressions for the cell loss probability of the PTBSF, PBSF and TBSF, assuming uniform traffic at the switch input. We refer the reader to the switches shown in Figure 2.

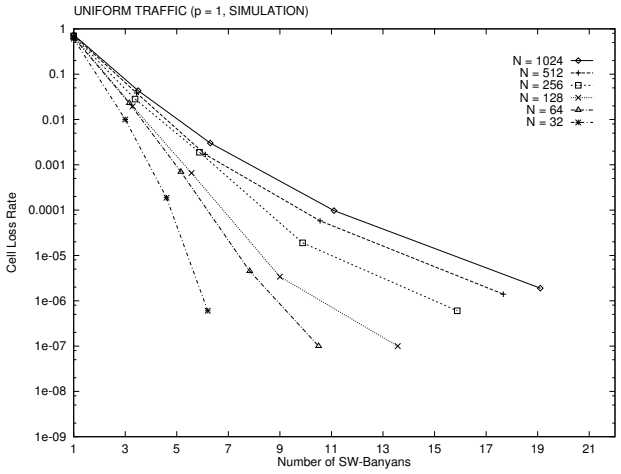

Figure 5: Loss rate in PTBSF for $N=1024$

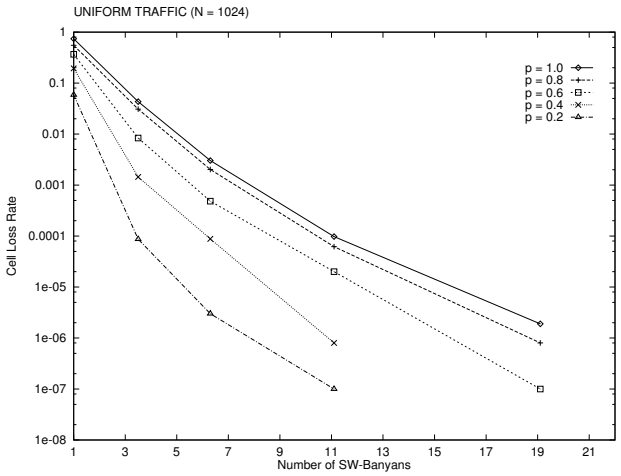

Figure 6: Loss rate in PTBSF for $N=1024$

For the case of PTBSF, assume a cell is issued into either inputs I_0 or I_1 of an SW-switch with a probability p, and that the probability of a cell on the vertical input I_v is q. Since the SW-switch has three inputs, we have three cases. First, one cell is present at inputs I_0 or I_1 with probability $p(1-p)(1-q)$ or at the vertical input I_v with probability $(1-p)^2q$. Second, two cells are present at I_0 and I_1 with probability $p^2(1-q)$ or at I_0 (or I_1) and I_v with the probability $p(1-p)q$. Third, three cells are present with probability p^2q.

The probabilities that a cell exits the switch at O_0, O_1, O_{v0}, and O_{v1} are given in Figure 3. This model allows propagating the pass-through probability in the horizontal direction and finding the probability of the cell being routed to the next lower level through the vertical outputs (see Figure 4). The probabilities on the vertical outputs at a particular level are used as inputs to the Banyans of the next lower level. The loss probability in the PTBSF is the sum of all probability of losses occurring at the outputs O_{v0} and O_{v1}

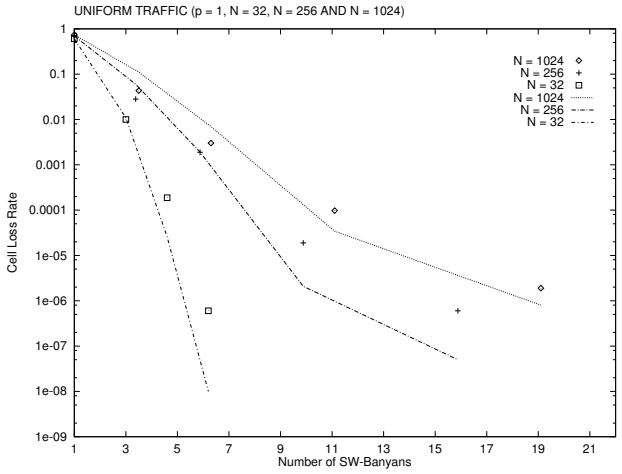

Figure 7: Loss rate in PTBSF at full load

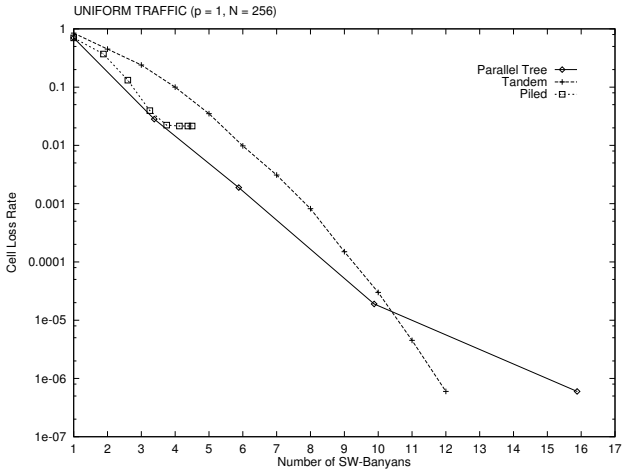

Figure 8: Loss rate with $N = 256$ and full load

of all SW-switches located in the lowest level of the PTBSF as shown in Figure 4.

The above analysis assumes that as the cells of the same time slot are propagated from one level/stage to the next, they remain independent. Hence, the effect of correlation is ignored. Simulation results indicate that the error introduced by this assumption becomes more significant with larger switches.

Figure 5 shows the cell loss rate obtained by simulation of the PTBSF as a function of the effective number of SW-Banyans for various values of N at full load ($p = 1$). A loss rate of 10^{-6} that is required for most ATM traffic sources is achievable for all the considered switch sizes ($32 \leq N \leq 1024$). To obtain a cell loss rate of 10^{-6}, the number of effective SW-Banyans should be 7 (5-level) for $N = 32$ and 20 (6-level) for $N = 1024$. Figure 6 shows the cell loss rate obtained by varying the load p for $N = 1024$. Figure 7 shows the cell loss rate results obtained from the analytical model (represented by lines) and from simulation (points) for the PTBSF as a function of the effective number of SW-Banyans and for some values of N at full load ($p = 1$). The analytical results deviate from the simulation results with increasing number of levels. The deviation is due to the fact that the analytical model ignores correlation among the cells.

The scalability of the PTBSF performance is another important feature. For example, with $N = 64$, the cell loss rate decreases from 10^{-4} to approximately 10^{-6} when the effective number of SW-Banyans increases from 7 to 9. The cell loss rate in the PBSF remains at about 10^{-2} regardless of hardware used.

Figure 8 shows that both TBSF and PTBSF exhibit acceptable cell loss rates (10^{-6} or lower) as required by most ATM traffic sources. In PTBSF, simple hardware is needed at the Banyan outputs because we only need to check whether an output carries a cell or not prior to routing it to the corresponding output buffer. However, the switching element in the PTBSF requires more hardware than that of the TBSF. Moreover, most of the complexity of the PTBSF is in the additional required wiring, namely, in the vertical cell forwarding.

Though both PTBSF and TBSF achieve a loss rate in the order of 10^{-6} or less in all studied cases, the PTBSF requires less hardware when the switch size is below $N = 128$. One can see that there is a crossover at $N = 256$ when comparing the required hardware cost for both PTBSF and TBSF to achieve a loss rate on the order of 10^{-6} as shown in Figure 8. TBSF requires less hardware than PTBSF for switch sizes of $N = 256$ and larger.

4 Performance under ATM traffic conditions

ATM networks are being engineered to support bursty and non-bursty sources with a wide range of bandwidth requirement. The communication services provided at the ATM layer consist of the following five service categories [**?**]: (1) Constant Bit Rate (CBR), (2) Real-Time Variable Bit Rate (rt-VBR), (3) Non-Real-Time Variable Bit Rate (nrt-VBR), (4) Unspecified Bit Rate (UBR), and (5) Available Bit Rate (ABR). The CBR and rt-VBR services are for real-time sources with hard cell delay and delay jitter requirements and limited tolerance to cell loss (ex: audio and video sources). At connection establishment, a source must negotiate a traffic contract specifying its Quality of Service requirements (QoS). For CBR sources, the only QoS parameter required is the Peak Cell Rate (PCR). For VBR, the PCR, the Sustainable (average) Cell Rate (SCR), and Maximum Burst Size (MBS) are declared. The nrt-VBR service is for applications with loose cell delay and delay jitter requirements and low cell loss (such as voice mail and some video applications). The UBR service category is a best effort service. Sources using this service are not required to specify any QoS parameters (ex: connection-less data). Similarly, for ABR traffic sources no traffic parameters are required. The main difference between UBR and ABR categories is that ABR traffic sources are congestion aware. That is, ABR sources make explicit use of congestion control mechanism to adjust their traffic based on feedback from the network. In contrast, UBR traffic is just discarded at congested switches with feedback given to the sources.

A realistic ATM workload is a mixture of bursty and

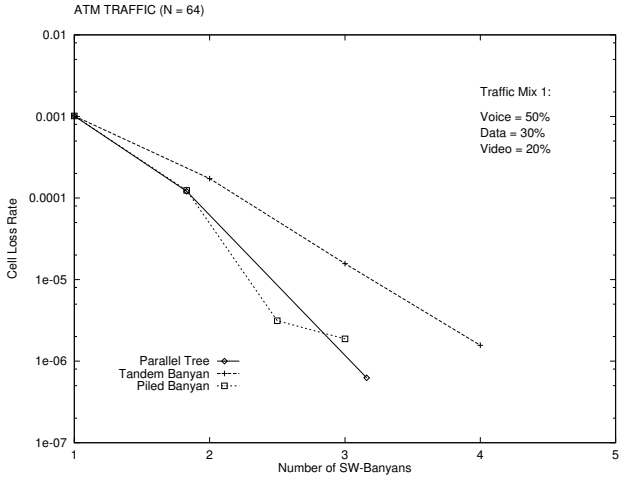

Figure 9: Loss rate for traffic 1

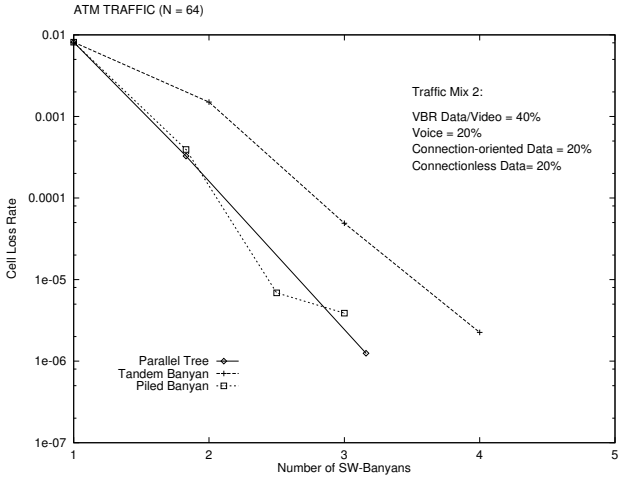

Figure 10: Loss rate for traffic 2

Type Source	Mean burst (cells)	Bit rate $SCR\times$ 384 *bps*	Burst-iness β	Cell loss rate
CBR	N/A	64 Kbps	1	10^{-4}-10^{-6}
CL	200	700 Kbps	up to 1000	10^{-12}
CO	200	25 Mbps	up to 1000	10^{-12}
VBR_v	2	25 Mbps	2 to 5	10^{-10}
B	3	1 Mbps	2 to 5	10^{-9}-10^{-10}
VBR_d	30	21 Mbps	2 to 5	10^{-9}

Table 1: Table

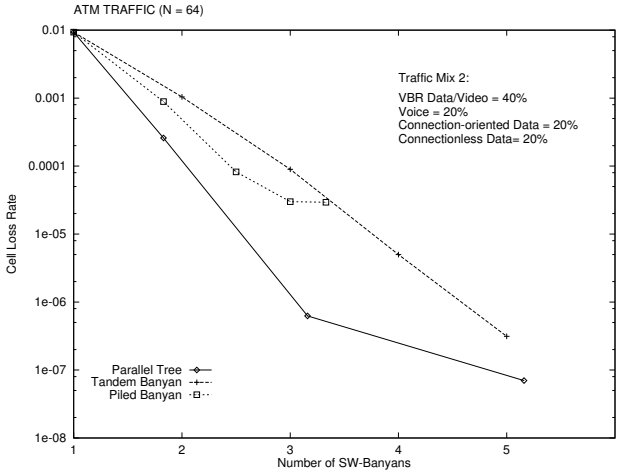

Figure 11: Loss rate for traffic 3

non-bursty sources with the load originated from a variety of traffic sources which exhibit correlation in space as well as in time. Traffic source characterization has been an extensive area of research[8]. A simple and widely adopted traffic source model is the *ON-OFF model*. According to this model, during the lifetime of a virtual connection, the traffic source will be in one of two states, *active* or *idle*. During the active state the source is transmitting cells at some given rate. Each active state may be followed by an idle period during which the source is silent. The cells generated during the same ON-period form a *burst*. Furthermore, it is always assumed that successive active and idle periods are statistically independent and exponentially distributed. As suggested by ITU-T, the length of the active period as well as that of the idle period are exponentially distributed.

For simulation purposes, several parameters have been identified, which together, completely characterize an ON-OFF traffic source. These are, the PCR, the SCR, and the average duration of the ON-state (t_{on}). Other parameters of interest such as the source burstiness (β) or the average duration of the OFF-state (t_{off}) are easily derived from these three parameters. For example, $\beta = \frac{PCR}{SCR}$ and $t_{off} = (\beta - 1)t_{on}$. Typical values for the traffic parameters for examples of traffic sources are summarized [8] in Table 1 which are: (1) CBR voice, (2) connectionless (CL) data, (3) connection oriented (CO) data, (4) VBR video (VBR_v), (5) Background (B) data and video, and (6) VBR video-data (VBR_d).

In our simulation study, we assumed that the PCR, t_{ON}, and β are known for each source. Furthermore, as recommended by ITU-T, we assumed that the active and idle periods are exponentially distributed with parameters $a = \frac{1}{t_{ON}}$ and $b = \frac{1}{t_{OFF}}$ respectively.

It is impractical to simulate the three Banyan switches for all possible traffic mixes. We experimented with the following traffic mixes. *Traffic Mix 1* is 20% Video (VBR), 50% Voice (CBR), and 30% Data (ABR and VBR). *Traffic Mix 2* is 40% of the sources are VBR Data/Video, 20% Voice (CBR), 20% Connectionless Data (ABR and VBR), and 20% Connection-Oriented Data (ABR). *Traffic Mix 3* is identical to *Traffic Mix 2* but with output concentration, where only odd (respectively even) numbered output ports are selected. *Traffic Mix 4* is identical to *Traffic Mix 3* but with output concentration, but this time only either the upper or lower output ports are selected. *Traffic-2* is a mix of higher percentage of burstier sources with larger bandwidth requirements than *Traffic-1*, i.e. much larger workload. This will expose the sensitivity of the switch to the workload. *Traffic-3* and *Traffic-4* generate as high a workload as *Traffic-2*, with the important difference that the level

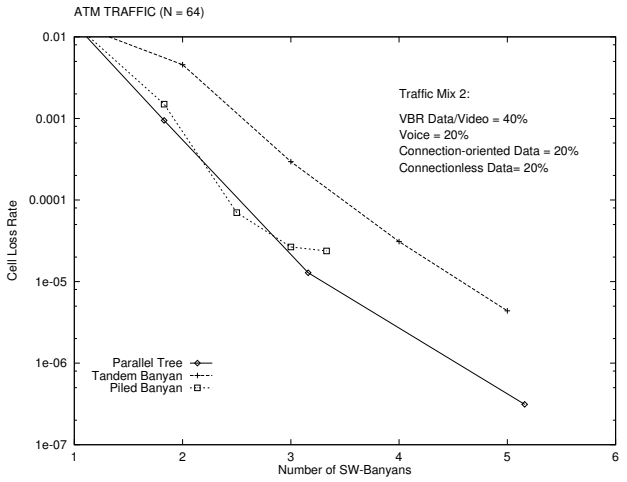

Figure 12: Loss rate for traffic 4

of internal conflicts as well as external conflicts is increased.

Figure 9 shows the loss rate versus the number of SW-Banyans for TBSF, PBSF, and PTBSF for N=64 using *Traffic-1*. The PTBSF switch generated the lowest loss rates compared to the other two switches. PBSF exhibited the poorest behavior, with the cell loss rate remaining at about 10^{-6} from level 3 and onward. We noticed also, that when N is increased, the performance of TBSF degrades much more rapidly than that of PTBSF. When N passed from 32 to 64, the cell loss rate has almost doubled for TBSF and was nearly the same for PTBSF and PBSF.

Figures 9-12 show the cell loss performance of TBSF, PBSF, and PTBSF for N=64 and for *Traffic-1, Traffic-2, Traffic-3*, and *Traffic-4* respectively. We observe that among the three switches, TBSF is the least sensitive to a change in the traffic mix. Both PBSF and PTBSF are sensitive to a change in the workload. However, we noticed that the sensitivity of PTBSF to the workload becomes less and less noticeable as we increase the number of levels (the effective number of SW-Banyans).

Figures 11 and 12 correspond to the traffic mix with odd/even and upper or lower output concentration respectively. Both traffic mixes generate more internal collisions than traffic with no output concentration. One may observe that for all three switches the cell loss performance for the lower/upper output concentration is noticeably worse than the case of odd/even output concentration. PTBSF still exhibits superior performance compared to TBSF or PBSF.

All figures show the cell loss rate as a function of the effective number of SW-Banyans varying from 1 to 5. With the PBSF switch, we always reach saturation for small values of the effective number of SW-Banyans (≤ 3). For the PTBSF switch, the simulator indicates a cell loss rate equal to zero when the effective number of SW-Banyans is greater than or equal to 5. For a TBSF switch with 5 levels, the cell loss rate varied between 5×10^{-6} (*Traffic-3/4*) and zero (*Traffic-1/2*).

5 Conclusion

We proposed a space-division ATM switch that consists of parallel Banyan structures arranged in a tree topology. Cell routing and contention resolution are done in a completely distributed manner without internal buffering, and there is no cell pre-processing prior to cell switching. It has been shown that scaling up the switch hardware provides as low cell loss rate as required. Though the vertical forwarding increases the switch complexity for VLSI implementation, the cell processing, delay jitter, and overall switching time are substantially lower compared to TBSF and PBSF. The architectural features are, guaranteed in-sequence delivery, low delay jitter, distributed control, self-routing, and fast switching. Experimental results indicate that the proposed switch has less cell loss rate and less switching delay than PBSF and TBSF under both ATM and uniform traffic, especially for small to medium size switches.

References

[1] Fouad A. Tobagi. Fast packet switch architectures for broadband integrated services digital network. *Proceedings of the IEEE*, 78(1):133–167, Jan. 1990.

[2] A. Saha and M. D. Wagh. Performance analysis of banyan networks based on buffers of various sizes. *IEEE Proc. INFOCOM'90*, 1:157–164, 1990.

[3] Reza Rooholamini, Vladimir Cherkassky, and Mark Garver. Finding the right ATM switch for the market. *IEEE Computer*, 27(4):17–28, Apr. 1994.

[4] J. S. Turner. New directions in communications (or which way to the information age?). *IEEE Commun. Mag.*, 24(10):8–15, Oct. 1986.

[5] P. C. Wong and M. S. Yeung. Design and analysis of a novel fast packet switch–Pipeline Banyan. *IEEE/ACM Transactions on Networking*, 3(1):63–69, Feb. 1995.

[6] Fouad A. Tobagi, Timothy Kwok, and Fabio M. Chiussi. Architecture, performance, and implementation of the tandem Banyan fast packet switch. *IEEE J. Selected Areas in Communications*, 9(8):1173–1193, Oct. 1991.

[7] Toshihiro Hanawa et al. Multistage interconnection networks with multiple outlets. *1994 International Conference on Parallel Processing*, I:1–8, 1994.

[8] G. D. Stamoulis, M. E. Anagnostou, and A. D. Georgantas. Traffic source models for ATM networks: a survey. *Computer Communications, Butterworth-Heinemann Ltd*, 17(6):428–438, June 1994.