Evaluation of Pipelined Dilated Banyan Switch Architectures for ATM Networks

Mayez Al-Mouhamed and Mohammad Kaleemuddin
Computer Engineering Department
CCSE, King Fahd University
Dhahran 31261, Saudi Arabia.
mayez,kaleem@ccse.kfupm.edu.sa

Abstract

In the *Pipeline Banyan* (PB) [1] the reservation cycle in the control plane is made several times faster than payload transmission in data plane. This enables pipelining multiple banyans. It is observed that the ratio of throughput to switching delay (service rate) is relatively low in the PB due to the banyan. For this, we present a scalable pipelined ATM switch architecture employing a family of *Dilated Banyan* (DB) networks together with their complexity analysis and performance. A DB can be engineered between two extremes: (1) a low-cost banyan with internal and external conflicts, or (2) a high-cost conflict-free fully-connected network with multiple outlets. Between the two extremes lies a family of DBs having different switching delays and throughputs. Increasing the dilation degree reduces path conflicts, which produces noticeable increase in service rate due to increase in throughput and decrease in path delay. Compared to PB, the *Pipelined Dilated Banyan* (PDB) requires less number of data planes for the same throughput, or provides higher throughput for a given number of data planes. Simulation of PDB was carried out under *uniform traffic* and simulated *ATM traffic*. We study performance under variation in the load, buffer size, and number of data planes. To analyse the robustness of the switch, we show that performance is not degradable under ATM traffic with temporal and spatial burstiness generated by using the ON-OFF model and some traffic mixes. The PDB is scalable with respect to service rate and can be engineered with respect to: (1) cell loss rate, (2) hardware resources, (3) size of buffers, (4) switching delays, and (5) delay incurred to higher priority traffic. The PDB can deliver up to 3.5 times the service rate of the PB with linear increase in hardware cost.

1 Introduction

Broadband integrated services digital networks (B-ISDN's) [2, 3, 4] are based on the use of *fiber optics* for transmission and *asynchronous transfer mode* (ATM) as the switching technology. To match the switching speed with the high transmission speed, ATM technology operates at extremely high speed with short fixed-length packets called cells. To gain efficiency and flexibility, ATM uses the connection-oriented approach [5] with statistical multiplexing to support a wide variety of data rates and rate-controlled traffic. Providing acceptable quality-of-service (QoS) for bursty real-time traffic requires minimal cell delay and delay jitter. No link-by-link flow control was planned to ensure minimal delays in buffering and switching. Only preventive actions against cell loss are provided by ATM. At connection set-up, resource is allocated based on currently available resource and active connections. For the above reasons, the *cell loss probability* (CLP) must be kept extremely low.

We are concerned with the design of a low complexity switch architecture capable of switching bursty traffic at typical rates on the order of 1 Gega cells per second, with a cell loss probability as low as 10^{-8}. The progress in the field of VLSI technology has brought new design concepts, high performance, high capacity, and low cost. Banyan-based multistage networks are suitable for VLSI implementation because of their modularity, self routing, and low hardware complexity. Several strategies have been suggested to overcome the banyan blocking feature.

In the *Tandem Banyan Switching Fabric* (TBSF) [6] the cells are issued to banyans arranged in series. One conflicting cell is misrouted in current banyan and re-issued to next banyan. Thus by properly adjusting the number of banyans the CLP can be made as low as needed. The cost is the relatively large number of banyans and implied propagation delays.

Multi-parallel banyans use vertical connections to shorten propagation delays such as in the *Piled Banyan Switching Fabric* (PBSF) [7] and *Parallel-Tree Switching Fabric* (PTBSF) [8] which have no input buffering. In the case of cell conflict some cells are vertically forwarded to a corresponding switching element in a parallel banyan to reduce the propagation delays due to different switching paths. However, achieving low CLP requires complex hardware and large number of connections.

The *Pipeline Banyan* (PB) [1] uses one single control plane for path reservation and a number of data planes for payload transfer. Each time slot consists of a number of reservation slots. In a reservation slot, a successfully self-routed header makes reservation of a path on a free data

plane to enable transfer of its payload. An input buffer may be notified to re-submit its cell header in next slot if its cell (header) cannot continue its self-route because of conflicts. Since the cell header in ATM is much shorter than payload, therefore, multiple reservation slots can be done during one payload transfer slot which enables pipelined operations. Thus, PB has low switching delay and relatively high throughput.

In this paper, we present an investigation of pipelined switch architectures employing a family of *Dilated Banyans* (DBs). Given the relatively low throughput and large switching delay of banyan our objective is to find an architecture that can provide higher throughput and less delays which enables increasing switch service rate. The DB was studied as packet switching in [9]. We show how the architecture of DBs can be made scalable with respect to throughput and propagation delays. We study pipelined switches employing few DBs for which we evaluate the: 1) switching delays, 2) number of data planes needed to guarantee some CLP, and 3) overall hardware complexity. This will allow us to find some pipelined schemes employing specifically designed DBs that can provide high throughput with reasonable switching delay.

The organization of this paper is as follows. Section 2 presents the background. Section 3 presents the topology and complexity analysis. Section 4 presents In section 5 we present evaluation under ATM traffic. In Section 6 we conclude this work.

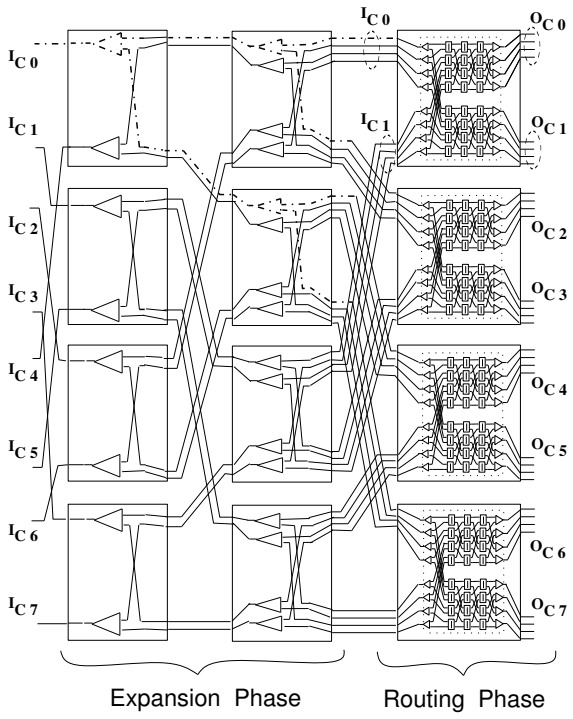

Figure 1: 1:4 Dilated Banyan with 8 inputs (each output channel has 4 ports)

2 Background

One fundamental problem to minimize CLP in ATM switches is to find an efficient method for partitioning the set of *head-of-line* (HOL) cells into subsets so that the cells within each subset are free of internal and external conflicts. Each subset of cells can then be switched out without conflicts by using a separate banyan. A number of proposals have been made to provide partial solutions. The idea is to issue all HOL cells to one banyan, perform self-routing, retrieve cells which reach their destinations, and re-issue all unsuccessful cells to the banyan, and so on. We call this approach *Iterative Conflict Resolution* (ICR) in which a cell loss occurs in: 1) last banyan in the case of no input buffering, or 2) input buffers in the case of full buffers.

In TBSF [6] cells are applied to banyans arranged in series. The TBSF [6] can achieve arbitrary low CLP at the cost of relatively long switching delay. Multi-parallel banyans use vertical connections among banyans to reduce sequential propagation delays. The PBSF [7] and PTBSF [8] are two examples. Hence, the throughput of the piled banyan saturate at 98% under full load, while the PTBSF can be scaled up to achieve arbitrary low CLP. Both switches have short propagation delay but they use relatively large amount of hardware and interconnections.

The PB [1] has one single control plane and a number of data planes. In a reservation slot, headers of all HOL cells are self-routed to their destinations within the control plane. In the case of conflict between two headers one of them is dropped and a back-pressure mechanism is used to notify the corresponding input buffer to re-submit its cell header in next slot. Multiple reservations can be done during payload transmission time. The reservation time is several times faster than payload transmission time which enables pipelined operations and contributes in reducing switching delay.

The PB can still guarantee very low CLP if overall cell arrival rate is below the service rate of the switch. We propose the design of a scalable banyan architecture that can provide higher throughput without dramatically increasing switching delays. Our objective is to investigate a class of dilated banyan networks that can be engineered between two extremes: (1) a low-cost banyan with internal and external conflicts, or (2) a high-cost conflict-free fully-connected network with multiple outlets. The idea is that reducing the degree of conflicts in dilated banyans is accompanied with shorter switching delays due to simplified hardware. For this reason, the service rate of *pipelined dilated banyan* (PDB) might largely exceed that of the PB.

3 Pipelined dilated banyans

We present a class of DBs for which the building blocks are: (1) 2×2 *binary sorter*, and (2) 1×2 *demultiplexer* (DM). Next we study the complexity of DBs through evaluation of the number of sorters and DMs, interconnection links, and propagation delays. Using a design approach for the sorters and DMs, we also find the complexity of needed hardware in terms of number of gates and express propagation delay as function of gate delay.

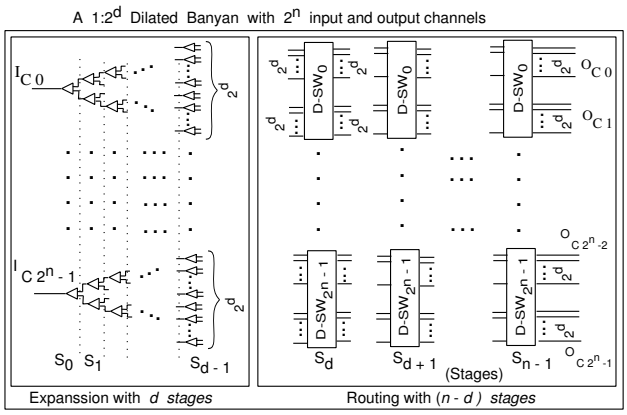

Figure 2: Components used the expansion and routing phases

3.1 A taxonomy of dilated banyans

The *Dilated Banyan* DB belongs to a class of dynamic, full access, single path, blocking multistage interconnection networks. The architecture of DBs is based on expanding the internal channel bandwidth (multiple ports) to reduce the CLP. A DB in which the degree of expansion of each link is d (1:2^d DB) is said to have 2^d-dilation. Figure 1 shows an 8-input 1:4 DB that has 3 stages and 8 output channels with up to 4 cells can be simultaneously transmitted over distinct links of a given channel. Each output channel has 4 ports (4-dilation). The dashed lines show one path through the binary expansion phase. The first two stages expand the input ports in the form of a binary tree by using DMs. The third stage is a routing phase for which the basic switch is called (D-SW) switch which has 2 input channels and two output channels. In the case shown in Figure 1 each channel of the D-SW switch has 4 ports. There is no cell loss in the first two stages. However, cell loss can occur in any *D-SW* of the routing phase. Here the routing phase consists of 4-stage binary sorters that sort the incoming cells based on their priority bit. This allows allocating the four output ports to most prior cells.

The DB has two phases: 1) *expansion phase*, and 2) *routing phase* (blocking). In the expansion phase the internal links are doubled at each of the d stages until reaching a 2^d-dilation at the dth stage. The kth stage of the expansion consists of $2^n \times 2^k$ DMs which are 1×2. Each DM routes an incoming cell to its upper or lower outputs depending on the cell destination bit.

The routing phase consists of $n - d$ stages, each contains 2^{n-1} switches called *D-SW* switches. Each *D-SW* switch has two input channels (I_{C0} and I_{C1}) and two output channels (O_{C0} and O_{C1}). Each of these channels has 2^d ports. Input and output channels are shown on Figure 1 for $n = 3$ and $d = 2$. At most 2^d cells with identical output can be simultaneously transmitted (one cell per port) without conflicts from input to output. A DB switch becomes non-blocking if the dilation degree is equal to the number of stages, i.e. there is no routing phase. In this case the hardware requirements are maximal due to its n-level 2^n binary trees. A non-blocking DB is one *Fully-Connected Network* with multiple outlets. On the other hand, a DB without an expansion phase is a simple banyan network (1:1 or 1-dilation) with internal and external conflicts.

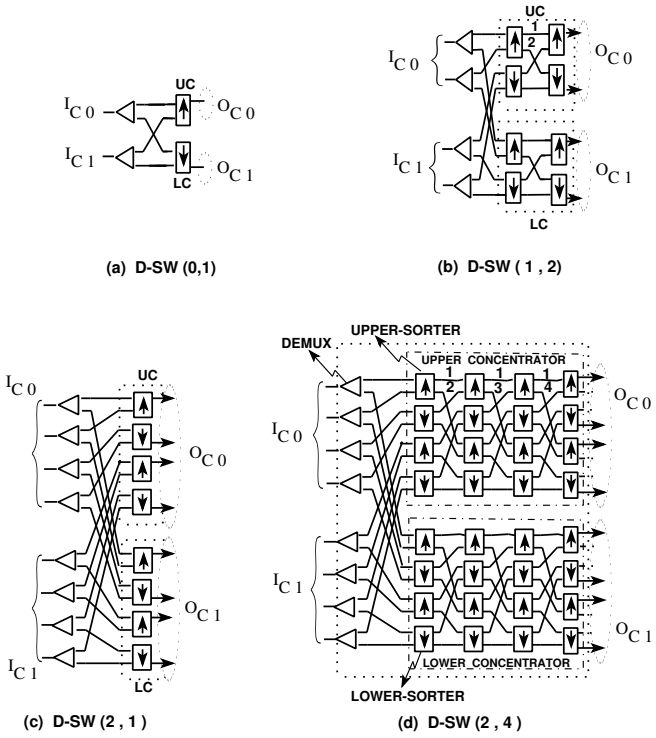

Figure 3: *D-SW*$(0, 1)$ (a), *D-SW*$(1, 2)$ (b), *D-SW*$(2, 1)$ (c), and *D-SW*$(2, 4)$ (d)

The *D-SW* has 2 input channels and 2 output channels where each channel has 2^d ports. Internally, the *D-SW* switch consists of one stage of DMs and two binary concentrators called *Upper Concentrator* (UC) and *Lower Concentrator* (LC). Each of the UC and LC may at most have 2^d stages as will be shown latter.

One may consider *D-SW* switches in which UC and LC are k-stage networks which we denote by *D-SW*(d, k). Simple banyans use *D-SW*$(0, 1)$, shown in Figure 3-(a), as a 2×2 switching element (SE). A 2-dilation DB uses *D-SW*$(1, 2)$, shown in Figure 3-(b), as basic SE. Figures 3-(c) and (d) show two *D-SW*$(2, k)$ for which $k = 1$ and $k = 4$, respectively. In *D-SW*$(2, 1)$ an internal conflict occurs at any of the one-stage sorters if there are two input cells at the sorter. In *D-SW*$(2, 4)$ a loss may occur only at any sorter of the last stage (fourth) if there are more than four cells at the inputs of the corresponding concentrator. If the DB is used as control plane in the pipelined scheme [1] then a cell that is one of the higher priority four cells succeeds in making path reservation in *D-SW*$(2, 4)$ and any of the other cells must be resubmitted again in the next cycle. Between *D-SW*$(2, 1)$ and *D-SW*$(2, 4)$ one can use other *D-SW*s ($k = 2$ and $k = 3$) with intermediate bandwidth. For example, in

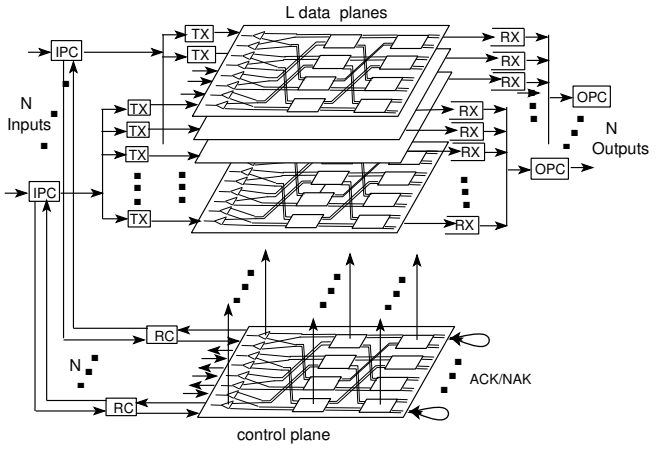

Figure 4: An 8×8 pipelined dilated banyan using DB(3,2,1)

D-SW$(2,2)$ a loss may occur if there are three cells on four input ports of any given channel.

For maximum throughput a 4-dilation DB may use *D-SW*$(2,4)$ as the basic SE, but *D-SW*$(2,1)$ or *D-SW*$(2,2)$ or *D-SW*$(2,3)$ could also be used to reduce cost and bandwidth. However, a cell that is one of the highest priority four cells is guaranteed to be allocated (succeeds) one output port of *D-SW*$(2,4)$. An n-stage, $1{:}2^d$ DB in which the routing phase is made of *D-SW*(d,k) is denoted by $DB(n,d,k)$, where $k \leq 2^d$. For example, the DBs shown in Figure 1 are identical to $DB(3,2,4)$.

The UC and LC have identical architecture and each has 2^{d+1} input ports and 2^d output ports. Depending on cell destination bit (x), the DM stage routes an incoming cell (up to 2^{d+1} cells) to one of the 2^d input ports of UC (for cells with $x = 0$) or to one of the 2^d input ports of LC (for cells with $x = 1$). Each of UC and LC sorts incoming cells (at most 2^{d+1} cells) based on cell priority bit and exit at most 2^d cells.

A *D-SW*$(n,d,2^d)$ outputs at most 2^d cells among the most prior cells. In this case, in each of UC and LC the lower priority cells in excess of 2^d cannot be correctly routed and must be resubmitted in the next reservation cycle. In other terms, each concentrator takes at most 2^{d+1} input cells, select at most 2^d cells among the most prior regardless of their source input port, and forward the selected cells to its 2^d output ports. Therefore, cell loss can occur within *D-SW* only if there are more than 2^d cells at the input of concentrator.

The UC and LC are made of stages of 2×2 binary sorters (up-sorters and down-sorters) interleaved with a perfect-shuffle permutation. If there are two input cells to an up-sorter (down-sorter), then the cell with highest priority exits at the upper (lower) output and the other cell exits at lower (upper) output. A single input cell is always sent to upper output regardless of its priority. The up-sorter and down-sorter have symmetric functions.

In the first sorter stage, the upper output of up-sorter transmits a cell if there is at least one cell at either inputs. The lower output transmits a cell if there are at least 2 input cells. In the ith sorter stage, the lower output of up-sorter transmits a cell if there is at least i cells at input of concentrator. In Figure 3-(d), we labeled some links with the minimum number of cells arriving on concentrator inputs so that at least one cell is transmitted on that link. Since the lower output has least priority, the lower outputs of ith stage up-sorter transmits a cell if and only if there is at least $i+1$ cells at inputs of concentrator. There are 2^{d+1} input ports and 2^d output ports for each of UC and LC. Therefore, there must be 2^d stages in each of the UC and LC to guarantee that no cell loss can occur in last stage of concentrator as long as the number of input cells is no more than 2^d at concentrator input. Moreover, a *D-SW*(n,d,k) for which $k = 2^d$ guarantees that 2^d cells among the most prior cells are successfully routed to the 2^d concentrator outputs among a set of 2^{d+1} input cells. It is shown that increasing the number of stages k in the concentrator produces: (1) increase in the throughput at output of concentrator, and (2) decrease in the probability of delaying higher priority cells.

3.2 Pipeline switch architecture by using dilated banyans

The architecture of an 8×8 pipelined dilated banyan PDB with 2-dilation is shown in Figure 4. The PSB has: (1) one dilated banyan control plane, (2) L dilated banyan data planes, (3) a set of input multiplexers, and (4) a set of output multiplexers. The vertical links of control plane provide the data planes with the state of all switching elements of control plane. Details about pipeline switch architecture can be found in [1].

3.3 Complexity analysis

In this section we evaluate hardware complexity and delays for a number of useful DBs which are *simple banyan* ($DB(n,0,1)$), *partially-dilated banyan* ($DB(n,d,1)$), and *fully-dilated banyan* ($DB(n,d,2^d)$). The concentrator of each *D-SW* consists of: (1) only one stage ($k = 1$) of sorters for simple banyan and partially-dilated banyan, and (2) 2^d stages of sorters for the fully-dilated banyan. The number of sorters, DMs, and links are shown in Tables 1 and 3.

We evaluate delays in control and data planes. There is n DMs along a path from input to output in a DB. A path through a *D-SW*(d,k) encounters k sorters, then the total delay due to sorters in the routing phase is $k(n-d)\tau_S$. Denote by τ^C_{DM}, τ^D_{DM}, τ^C_S, and τ^D_S the delays through a DM and a sorter in control (C) and data (D) planes, respectively. The overall delay in DB due to expansion and routing phases is then $DB^C_\tau(n,d,k) = n\tau^C_{DM} + k(n-d)\tau^C_S$ in control plane and $DB^D_\tau(n,d,k) = n\tau^D_{DM} + k(n-d)\tau^D_S$ in data plane. Note that paths in data plane are pure combinational logic.

A sorter requires three D-latches to store the activity bits of two cells and the state, state logic, and two parallel 2×1 multiplexers. Using dual input gates, 12 gates are needed for 3 D-latches. The delay of one D-latch is $2\tau_{gate}$, where τ_{gate} is one gate delay. The state is $pr_0 a_0 + \overline{a_1}$, where pr_0, a_0, and a_1 are the priority of upper input cell (C_0), activity of C_0, and activity of lower input cell, respectively. The state requires 3 gates and has a delay of $2\tau_{gate}$. The two

	Banyan $BD(n,0,1)$	Partially-dilated $BD(n,d,1)$
$DB_{Sorters}$	$n2^n$	$(n-d)2^{n+d}$
DB_{Dmux}	$n2^n$	$(n-d+1)2^{n+d}-2^n$
DB_{Link}	$(3n+1)2^n$	$[3(n-d)+2]2^{n+d}-2^n$
$\tau_{control}$	$14n\tau_{gate}$	$(14n-8d)\tau_{gate}$
τ_{data}	$3n\tau_{gate}$	$(3n-2d)\tau_{gate}$
Complex. (Control)	$32n2^n$	$32(n-d)2^{n+d}+$ $11\times 2^n(2^d-1)$
Complex. (Data)	$16n2^n$	$16(n-d)2^{n+d}+$ $6\times 2^n(2^d-1)$

Table 1: Hardware requirments in partially dilated banyans

	Fully-Dilated $BD(n,d,2^d)$
$DB_{Sorters}$	$(n-d)2^{n+2d}$
DB_{Dmux}	$(n-d+1)2^{n+d}-2^n$
DB_{Link}	$[(n-d)(2^{d+1}+1)+2]2^{n+d}-2^n$
$\tau_{control}$	$[6n+8(n-d)2^d]\tau_{gate}$
τ_{data}	$[n+2(n-d)2^d]\tau_{gate}$
Complexity (Control Plane)	$(21\times 2^d+11)(n-d)2^{n+d}$ $11\times 2^n(2^d-1)$
Complex. (Data)	$(10\times 2^d+6)(n-d)2^{n+d}$ $6\times 2^n(2^d-1)$

Table 2: Hardware requirments in fully dilated banyans

multiplexers require 6 gates and each has a delay of $2\tau_{gate}$. Hence a sorter in control plane requires $N_S^C = 21$ gates and has a delay of $\tau_S^C = 8\tau_{gate}$. In data plane the sorter is reduced to two 2×1 multiplexers that are set by using one D-latch. Thus a sorter in data plane requires $N_S^D = 10$ gates and has a delay of $\tau_S^D = 2\tau_{gate}$.

A DM requires two D-latches to store the activity bit of input cell and the state, state logic (up or down), and one 1×2 demultiplexer. The state is x_0a_0, where x_0 is destination. Hence a DM in control plane requires $N_{DM}^C = 11$ gates and has a delay of $\tau_{DM}^C = 6\tau_{gate}$. In data plane the DM is reduced to the 1×2 demultiplexer that is set by using one D-latch. Thus a DM in data plane requires $N_{DM}^D = 6$ gates and has a delay of $\tau_S^D = \tau_{gate}$.

Tables 1 and 3 list the delays $\tau_{control}$ and τ_{data} of dilated banyans as function of gate delay. The total complexity of DB is evaluated by using the expression of $DB_{sorters}$ and DB_{Dmux} and the values of N_S^C, N_{DM}^C, N_S^D, and N_{DM}^D.

4 Evaluation under uniform traffic

In pipelined switches the cells are generated in the beginning of each time slot which consists of a number of reservation slots. HOL cells are submitted to control plane in each reservation slot. In unbuffered switches only cells generated during the current time slot are submitted to the control plane for each reservation slot until they succeed in reserving a path or being considered as lost.

4.1 Performance of pipelined fully-dilated banyans

The *Pipelined Fully-Dilated Banyan* (PFDB) uses *fully-dilated banyan* $DB(n,d,2^d)$ having 2^d-sorter stages in each *D-SW* switch of the routing phase. The throughput of $DB(n,d,2^d)$ is the highest among the family of DBs.

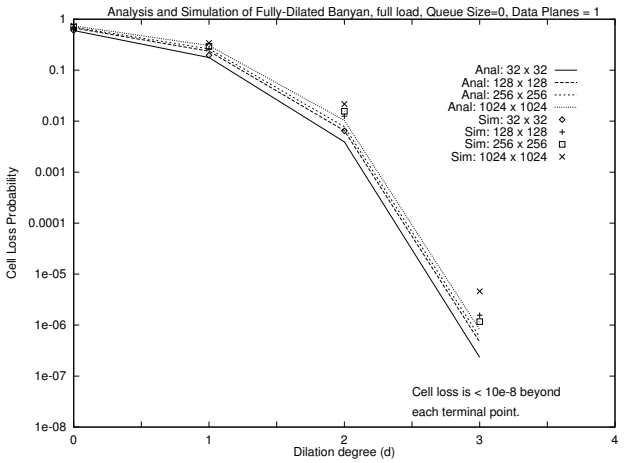

Figure 5: CLP of fully-dilated banyans ver. dilation degree

This means we use one $DB(n,d,2^d)$ in control plane and in each data plane of the pipelined switch. Figure 5 shows the CLP of single $DB(n,d,2^d)$ plane versus the dilation degree d as obtained from analysis and simulation.

Here a dilation degree of 3 (8-dilated) is sufficient to achieve a CLP around 10^{-6} at full load but at a high cost of the hardware complexity compared to that of simple banyan. This is discussed latter in the hardware analysis of Sub-Section 4.5. Note that fully-dilated banyans exhibits nearly the same CLP regardless of the switch size.

The analytical results were more optimistic than simulation results because the analytical model assumes uniform traffic at all the stages. In reality the uniform traffic gets corrupted by the deterministic stage routing which increases from one stage to the next. In other terms, the traffic uniformity decreases with increasing stage number. We validated the above interpretations by uniformly re-generating the cell destination after each stage of the simulation which gave CLPs that were very close to those obtained from the analysis. Therefore, the simulation results are more representative of real switch performance than the analytical results. We also evaluated the analytical model of pipelined DBs by using our analytical model of $DB(n,d,2^d)$ and the *discrete-time probability state transition* used in [1] to model the number of backlogged cells in input queues. The CLP obtained from analysis was always more optimistic than that obtained from simulation due to the effects described above.

Buffering of single plane DB did not produce significant drop in CLP except when the dilation degree was 3 or above, i.e. when CLP is very low. For the pipelined scheme we used fully-dilated banyan with dilation degree 1 (2-dilation) and observed good results. As shown on Figure 6 an unbuffered PFDB requires only 5 or 6 data planes as compared to 12 data planes for an unbuffered PB (see Figure 10). The use of input buffering significantly reduces the number of data planes (reservation cycles) needed to achieve some CLP. From Figure 6, a 256-input unbuffered PFDB requires 5 data planes to achieve a CLP below 10^{-6}, while the same

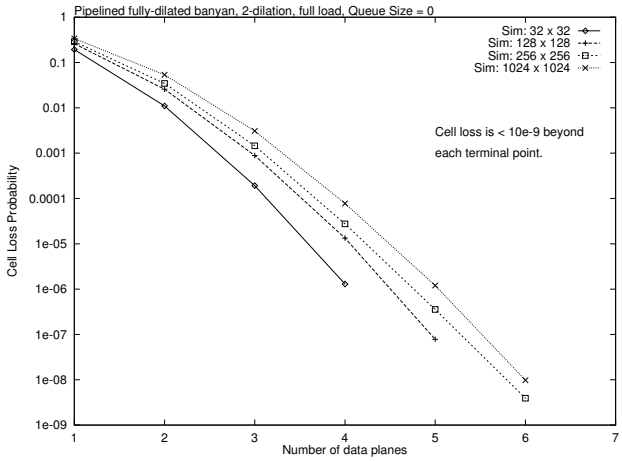

Figure 6: CLP of unbuffered pipelined fully-dilated banyans

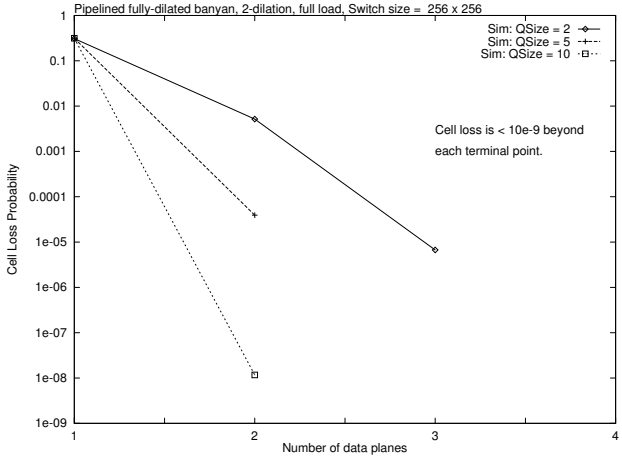

Figure 7: CLP of buffered pipelined fully-dilated banyans

level of performance can be achieved by using only 4 data planes if 2 buffers were used. When input buffering is used, the number of data planes needed to achieve a CLP of nearly 10^{-8} gradually decreases from 4 to 2 with increasing the buffer size from 2 to 10 as shown on Figure 7 for the case of a switch size of 256. This Figure shows the profitability of increasing input buffer size on a 256×256 PFDB.

4.2 Performance of pipelined partially-dilated banyans

The *Pipelined Partially-Dilated Banyan* (PPDB) uses one $DB(n, d, 1)$ in each data plane. We have simulated the $PPDB$ for $d = 1$ (2-dilation) and $d = 2$ (4-dilation). Figures 8 and 9 show the CLP of $PPDB$ in the case of 4-dilation. With no input buffering the 4-dilated $PPDB$ requires 6 or 7 data planes (8 or 9 data planes for 2-dilation) against 12 for the PB (See Figure 10).

The CLP reduces further when input buffers are used. We observe a similar trend as in unbuffered pipelined switches. When a CLP below 10^{-6} is needed, the profitability of input buffering is significant as it may reduce the number of data planes from 3 or 4 for 2-dilation to 2 for 4-dilation

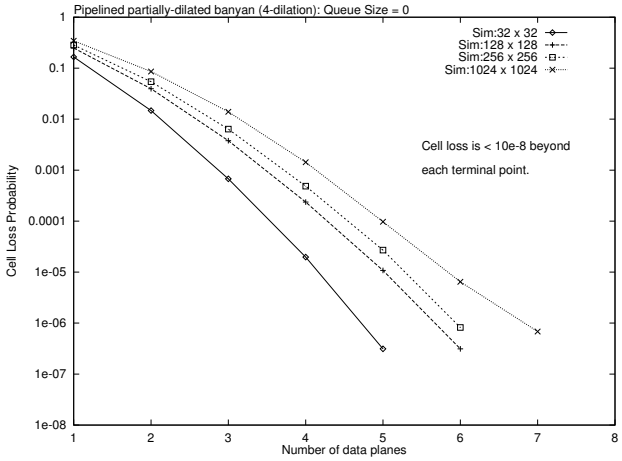

Figure 8: CLP of unbuffered pipelined partially-dilated (4) banyans

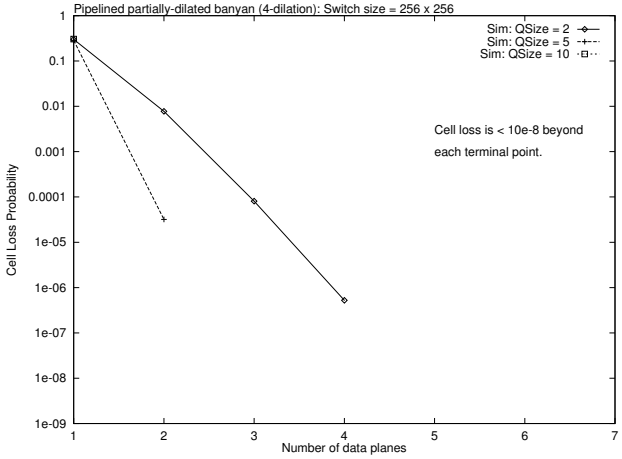

Figure 9: CLP of buffered pipelined partially-dilated (4) banyans

(Figure 9). Notice that switching delay of partially-dilated banyan is shorter than that of simple banyan because of the expansion phase. Therefore, the use of partially-dilated banyans allows reducing the number of data planes in the pipelined switch while decreasing the reservation time compared to that of simple banyan. This reduction in the number of data planes and delay in pipelined partially-dilated banyan is gained at the cost of increased hardware and number of interconnection links in both control and data planes. Analysis of gained performance and cost will be presented at the end of this section.

4.3 Performance of pipelined banyans

The *pipelined banyan* (PB) switch refers to the scheme presented in [1]. Notice that the simple banyan network is a particular DB ($DB(n, 0, 1)$) for which the dilation degree $d = 0$ and the number of sorter stages in each D-SW is 1.

In unbuffered PB the loss can occur in the last reservation slot, while in buffered PB the loss occurs at input of full buffers. Figure 10 shows the CLP as function of the

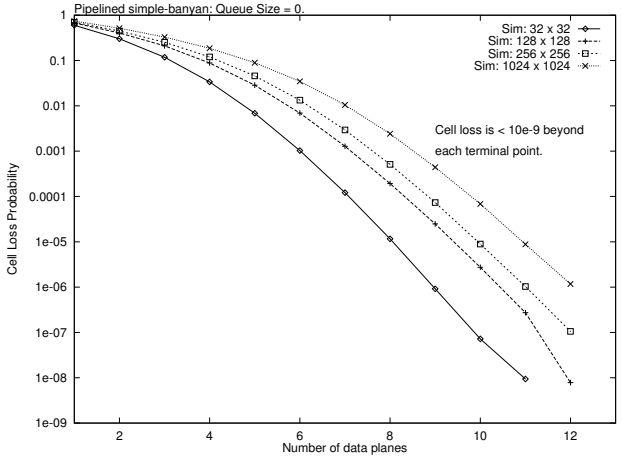

Figure 10: CLP of unbuffered pipelined banyans

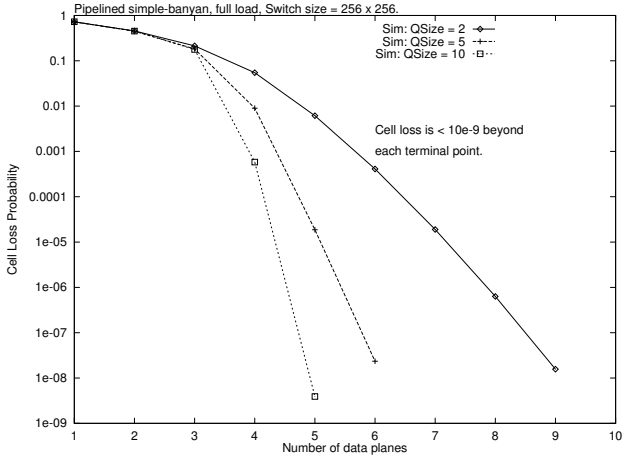

Figure 11: CLP of buffered pipelined banyans

number of data plane for unbuffered PB. To achieve a CLP below 10^{-6} the number of data planes required is 12 for large switches such as 256×256 or 1024×1024. The number of data planes required can be reduced when input buffers are available. In buffered PB, a cell that fails in a reservation slot remains in input buffer until it succeeds in some subsequent reservation slot. Figure 11 shows the profitability of input buffering for pipelined simple banyan. This Figure shows the effect of varying input buffer size for a 256×256 switch. When the input buffer size is increased from 2 to 10 we observe a decrease in the number of data planes from 9 to 5 while achieving a CLP around 10^{-8}.

Though the simple banyan has relatively poor throughput, input buffering contributes significantly in boosting the performance of the pipeline banyan. Our simulation indicates that six or seven data planes are required to achieve a CLP below 10^{-6} for large switches with 10-input buffering.

4.4 Queuing delays

In this section we study total switch delay which the sum of *HOL delay* and *queuing delay*. It is expressed in reservation cycles. The HOL delay is counted from the time a cell becomes HOL to the time the cell succeeds in making a reservation. The queuing delay is counted from the time a cell enters some input buffers until the time at which the cell becomes HOL.

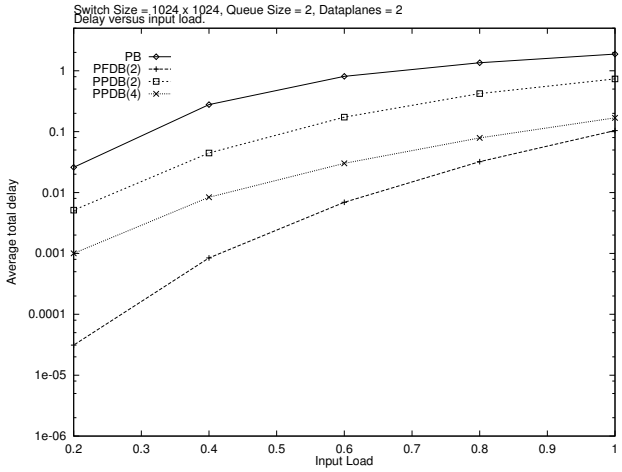

Figure 12: Comparison of average total delay

Figure 12 shows the average total delay for PB, $PFDB$ with 2-dilation, $PPDB$ with 2-dilation, and $PPDB$ with 4-dilation when 2 input buffers are available for each input port. Lighter input loads certainly contributes in shortening the HOL delay and total delay for all switches. Mainly, delays are important when input traffic is close to full load. Pipelined switches employing 2-dilation $PFDB$ or 4-dilation $PPDB$ have high throughput which indicates that a cell remains, on the average, HOL for small fraction (about 10^{-2}) of a reservation cycle at full load. At full load, the throughput of PB and $PPDB$ (2-dilation) is much less compared to the other switches which causes an HOL delay of 10^{-1}. The total delay (Figure 12) for PB and $PPDB(2)$ with the same conditions becomes closer to 1 reservation cycle. In the case of PB and $PPDB(2)$ at full load, a cell may: 1) waits in input buffer for 1 clock before becoming HOL, or 2) finds the buffer empty and succeeds in the reservation attempt at second cycle. It is observed that HOL delay dominates the total delay in the case of 2-input buffers. Therefore, input buffering is profitable and buffer size significantly affects CLP.

4.5 Comparisons

The performance of pipelined switches cannot be characterized only by the achieved throughput or the corresponding CLP. The time it takes to achieve a level of performance is one important factor in the evaluation. For example, the pipelined banyan can achieve a CLP below 10^{-6} when there are six data planes. Whether this performance is adequate for ATM rates or not depends on: (1) the number of needed reservation slots, (2) switching delay in control plane, and (3) payload transmission time in data plane.

Using the results from Section 3.3 for a 256-input dilated banyan switch we list the hardware requirements in Table 3. To compare performance of the pipelined switches that em-

	PB	PDB_2	PDB_4	PDB_8	FDB_2
$DB_{Sorters}$	2048	3584	6144	10240	7168
DB_{Dmux}	2048	3840	6912	12032	3840
DB_{Link}	6400	11520	20224	34560	18688
$\tau_{control}$	112	104	96	88	160
τ_{data}	24	22	20	18	36
Compl. in C.P.	65536	117504	205056	347392	192768
Comp. in D.P.	32768	58880	102912	174592	94720

Table 3: Hardware requirements and delays for 256×256 dilated banyans

ploy DBs we need to evaluate the service rate as function of gate delays. The service rate of the switch is the number of cells that can be switched per unit of time. Though wires occupy significant space and cause complexity in VLSI and PCB we do not account for wiring and wire delays for simplicity. This assumes that delays through a logical gate in VLSI dominate delays on wire.

A payload of N_{cell} bits can be transmitted over a free data plane in $T_{data} = N_{cell}\tau_{bit} + \tau_{data}$ gate delays, where τ_{data} is the delay in one data plane and τ_{bit} ($1/\tau_{bit}$ is the rate) is the time in gate delays to remove one bit from input buffer. In the control plane, the switching time (also reservation slot) of the header is $T_{control} = \tau_{control}$ gate delays which accounts for self-routing of the header along a path from input to output. Note that $N_{dp} = \lceil T_{data}/T_{control} \rceil$ is the least number of needed data planes in the pipelined switch to guarantee there is a free data plane for payload transmission at the end of each reservation slot.

Denote by p the cell input load and let clp be the switch CLP when each time slot is formed by L reservation slots. In steady state, the average number of cells that can be switched in one time slot is $p(1-clp)2^n$ for 2^n-input switch. For pipelined switches each time slot consists of L reservation slots but this requires $N_{dp}^{phys} = Min\{N_{dp}, L\}$ physical data planes. The duration of one time slot is then $T = L\tau_{control}$ gate delays. The time to switch $p(1-clp)2^n$ cells is T gate delays. If one assumes N_{dp}^{Phys} data planes, then the service rate (S) of the pipelined switch will be:

$$S = \frac{p(1-clp)2^n}{L\tau_{control}}$$

We assume header and payload are stored into the input buffers and retrieved at a rate of *one bit per gate delay* ($\tau_{bit} = 1$). One gate delay is assumed to be 1 ns. Now we consider a number of buffered pipelined DBs which are PB, $PPDB_2$, $PPDB_4$, and $PFDB_2$ and list some of their parameters in Table 4. L is being the number of reservation slots required to achieve a CLP below 10^{-8}.

For PB the number of needed reservation slots is $L = 6$. Since $N_{dp} = 4$ we can make 6 reservation cycles by using 4 data planes. The same considerations apply to the other switches. Due to low throughput of simple banyan, PB requires higher number of data planes than a pipelined DBs which requires between 2 and 3 more complex data planes.

The service rate is evaluated for PB, $PPDB_2$, $PPDB_4$, and $PFDB_2$ with 256 inputs and outputs. Since the number of reservation slots L is dictated by the need for an

Buffered pipelined dilated banyan (input buffer size is 10)

	PB	$PPDB_2$	$PPDB_4$	$PFDB_2$
Res. slots L	6	3	2	2
Achieved CLP	$4\times$ $\times 10^{-9}$	$1.2\times$ $\times 10^{-8}$	$2\times$ $\times 10^{-8}$	$1.2\times$ $\times 10^{-8}$
T_{data}	448	446	444	460
$T_{control}$	112	104	96	160
N_{dp}	4	5	5	3
N_{dp}^{Phys}	4	3	2	2
Control plane	65536	117504	205056	192768
Data planes	131072	176640	205824	189440
Total Links	32000	46080	60672	56064
Total Gates	196608	294144	410880	382208
Service Rate (Gega Cells/s)	0.381	0.821	1.333	0.800

Table 4: Service rates of 256×256 pipelined dilated banyans

acceptable CLP. Pipelining allows minimizing the cost of L reservation slots through the use of only N_{dp}^{Phys} data planes. Therefore, the service rate S is one *structural feature* of the used banyan which cannot be increased beyond some limit for a given design technology. The service rate of PB is the least for the family of pipelined DBs. According to our design model and assumptions, a 256-input PB can switch cells with a CLP below 10^{-8} only when overall cell arrival rate is below 381×10^6 cells/s. This represents a structural limit of pipelined banyans.

One way to achieve higher service rates than that of PB is to use DBs. The highest service rate is achieved for $PPDB_4$ which indicates that a 4-dilation with one-stage of sorters in the *D-SW* switch was critical in providing high throughput without dramatically increasing the reservation time. It is shown in Table 4 that a $PPDB_4$ has nearly 3.5 times the service rate of PB at hardware cost of nearly three times that of PB. There are two reasons for the relatively high service rates of partially-dilated banyans. First, the non-blocking binary expansion phase is responsible for significantly increasing the rate of successful reservations compared to simple banyans. Second, the use of one sorter stage in each *D-SW* switch of partially dilated banyans was the key factor to maintain a low propagation delay compared to fully-dilated banyans.

The number of stages in the *D-SW* controls the probability of delaying higher priority traffic. Increasing the number of stages in *D-SW* causes a decrease in the probability of delaying higher priority traffic. Although $PPDB$ generally provides higher service rate, the $PFDB$ incurs the least delays to higher priority traffic.

5 Performance under ATM traffic conditions

A realistic ATM workload is a mixture of bursty and 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[10]. 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

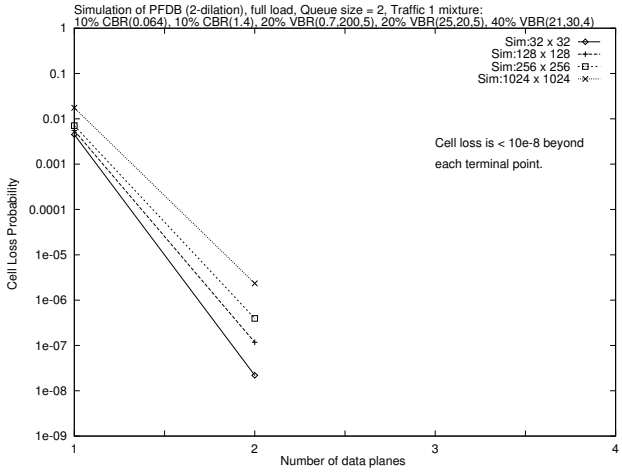

Figure 13: PFDB with 2-dilation under ATM *Traffic-1*

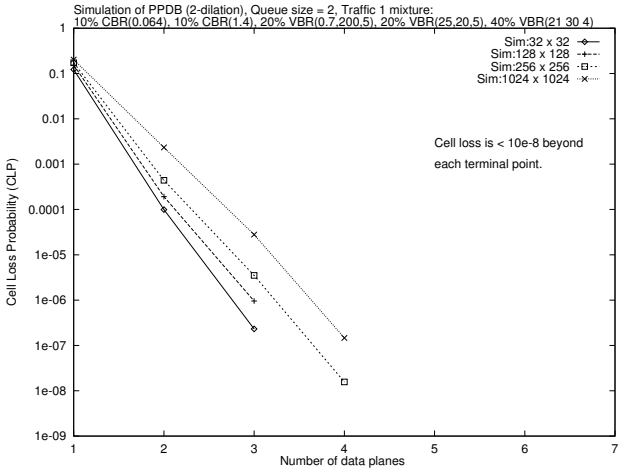

Figure 14: PPDB with 2-dilation under ATM *Traffic-1*

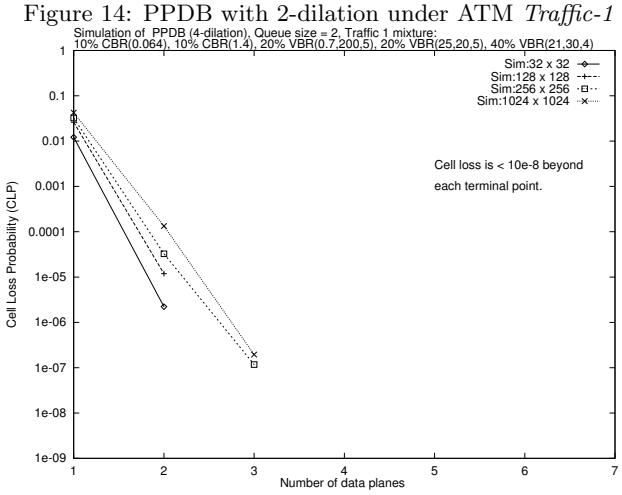

Figure 15: PPDB with 4-dilation under ATM *Traffic-1*

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.

For simulation purposes, several parameters have been identified, which together, completely characterize an ON-OFF traffic source. These are, the peak cell rate (pcr), the sustainable cell rate (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, $\beta = pcr/scr$ and $t_{off} = (\beta - 1)t_{on}$. Typical values for the traffic parameters for some traffic sources are summarized in [10].

In our simulation study, we assumed that pcr, t_{on}, and β are known for each source. Furthermore, we assumed that the active and idle periods are exponentially distributed with parameters $a = 1/t_{on}$ and $b = 1/t_{off}$, respectively.

We subjected the pipelined banyan PB, pipelined partially-dilated banyan $PPDB$, and pipelined fully-dilated banyan $PFDB$ switches to *Traffic 1* and *Traffic 2*. For *Traffic-1*, it was observed that less number of data planes were required to achieve some CLP than that needed in the case of uniform traffic for equal number of input buffers. The above observation was true for all three switches and for all switch sizes. We experimented with the following traffic mixes. *Traffic Mix 1* consists of: 10% CBR(pcr=0.064), 10% CBR(1.4), 20% VBR(scr=0.7,$t_{on}(in\ cells)$ =200,β =5), 20% VBR(25,20,5), and 40% VBR(21,30,4). *Traffic Mix 2* consists of: 25% CBR(0.064), 25% CBR(1.4), 12% VBR(0.7,200 5), 13% VBR(20,25,5), and 6% VBR(2,25,10).

The reason is that *Traffic-1* and *Traffic-2* causes less conflicts compared to full load uniform traffic. Figure 13 shows the CLP for $PFDB$ switch with 2-dilation and 2 input buffers under *Traffic-1*. Two data planes were required to achieve a CLP close to 10^{-6} for $PFDB$ and between 3 and 4 for $PPDB$ as shown on Figure 14. Using the same number of data planes, the CLP of $PFDB$ dropped below 10^{-8} for all switch sizes when the buffer size was increased to 5.

We repeated the above experiments by using *Traffic-2* with 2 input buffers and noticed that the CLP is much less than that obtained under *Traffic-1* for the same number of data planes. Specifically, the CLP of PB, $PPDB$, and $PFDB$ was below 10^{-8} when 2 data planes and two or more input buffers were used with all switch sizes. The reason is that in *Traffic-2* only 50% of the sources are VBR sources compared to 80% in the case of *Traffic-1*.

Figure 15 shows the CLP for $PPDB$ with 4-dilation when 2 input buffers are used under *Traffic-1* which can be compared to the case of 2-dilation shown on Figure 14. Increasing the dilation degree of $PPDB$ under *Traffic-1* leads to lower CLP or lesser number of needed data planes which indicates that CLP is significantly affected by the degree of dilation under both uniform traffic and ATM traffic. Figure 16 shows the effect of varying the input buffer size for $PPDB$ with 2-dilation under *Traffic-1*.

Figures 17 and 18 show the CLP for PB when using 2 input buffers under *Traffic-1* and *Traffic-2*, respectively. In the case of *Traffic-2*, much less number of data planes are

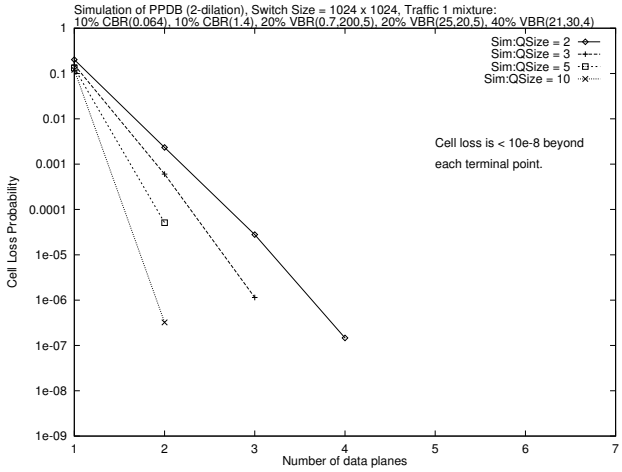

Figure 16: Increasing buffer size on 1024-input PPDB(2) under *Traffic-1*

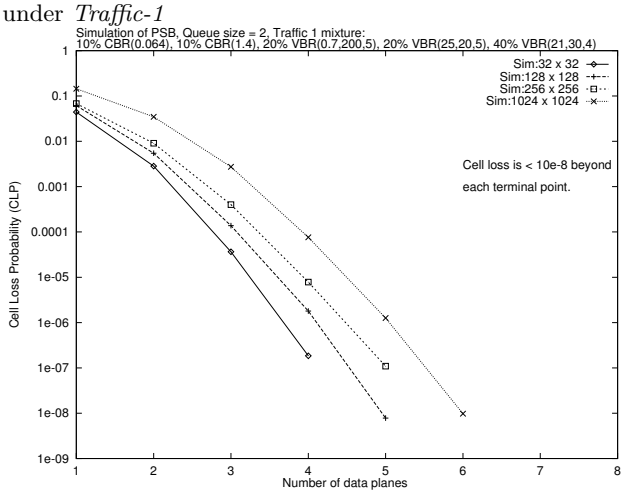

Figure 17: CLP of PB with 2 input buffers under *Traffic-1*

needed to achieve equal CLP which indicates that PB is also very sensitive to percentage of VBR sources. Using other experiments, we find that increasing the buffer size can be rewarded by a significant drop in the number of data planes needed to achieve a CLP of 10^{-6} or less. For example, the number of data planes dropped from 6 to 4 when increasing input buffering from 2 to 10 in the case of a 1024-input PB.

6 Conclusion

In this paper we evaluated pipelined ATM architectures employing a family of *Dilated Banyans* DBs for increasing throughput and reducing switching delay in banyan-based ATM switches. A DB can be engineered between two extremes: (1) a low-cost banyan with internal and external conflicts, or (2) a high-cost conflict-free fully-connected network with multiple outlets. Between the two extremes lies a family of DBs having different switching delays and throughput. Increasing the dilation degree reduces path conflicts which produces noticeable increase in service rate due to increase in throughput and decrease in path delay. Dilated

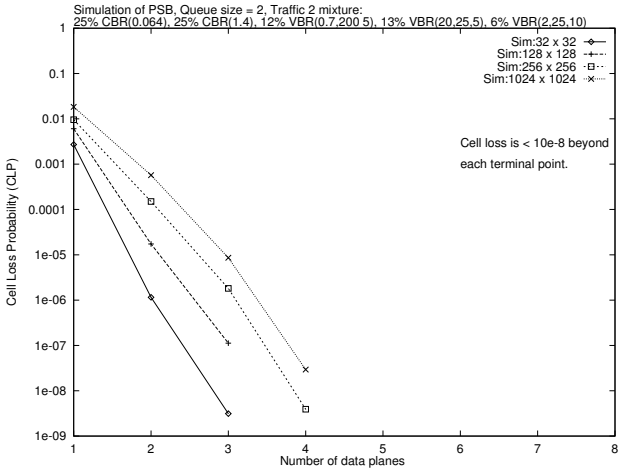

Figure 18: CLP of PB with 2 input buffers under *Traffic-2*

banyans were studied with respect to their hardware requirements, number of interconnections, input buffering, switching delay, and CLP. We carried out complexity analysis and simulation of pipelined dilated banyans which we subject to *uniform traffic* and *ATM traffic*. We also studied the robustness of the proposed pipelined switches under a variety of ATM traffic. Pipelining partially-dilated banyan can provide up to 3.5 times the service rate of pipeline banyan with linear increase in hardware cost.

7 Acknowledgment

We acknowledge contribution from Dr. Habib Youssef from the Computer Engineering Department, King Fahd University of Petroleum and Minerals (KFUPM), in the implementation of the well known ON-OFF traffic model. The authors acknowledge computing support and conference attending support from King Fahd University of Petroleum and Minerals, Dhahran 31261, Saudi Arabia.

References

[1] 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.

[2] M. Kawarasaki and B. Jabbari. B-ISDN architecture and protocol. *IEEE J. Selected Areas in Communications*, 9(9):1405–1415, Dec. 1991.

[3] D. Delisle and L. Pelamourgues. B-ISDN and how it works. *IEEE Spectrum*, 28(8):39–42, Aug. 1991.

[4] Martin de Prycker. *Asynchronous Transfer Mode - solution for broadband ISDN*. Ellis Horwood, 1991.

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

[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] M. Al-Mouhamed, H. Yousef, and W. Hassan. A Parallel-Tree Switch Architecture for ATM networks. *Inter. Journal of Communication Systems*, Vol 11, No 1, 1997.

[9] T. T. Lee and S. C. Lieu. Broadband packet switches based on dilated interconnection networks. *IEEE Trans. on Communications*, Vol 42(2/3/4):732–744, 1994.

[10] 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, Jun. 1994.