DESIGN AND EVALUATION OF A HIGH PERFORMANCE MULTI-PRIORITY MULTICAST ATM AND IP SWITCH

by

JOSEPH CHU

B.A.Sc., Queen's University, Canada, 1996

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENT FOR THE DEGREE OF MASTER OF APPLIED SCIENCE

IN

THE FACULTY OF GRADUATE STUDIES

DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

We accept this thesis as conforming

to the required standard

THE UNIVERSITY OF BRITISH COLUMBIA

June, 2002

In presenting this thesis in partial fulfilment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the head of my department or by his or her representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Department of Electrical and Computer Engineering.

The University of British Columbia
Vancouver, Canada

Date June 3/2002,

Abstract

Asynchronous Transfer Mode (ATM) and Internal Protocol (IP) are two commonly used protocols for the demands of high speed networking technology. The switching technologies employed in ATM cell and IP packet switches have seen extensively researched and studied in recent years. However, most of the switches developed have room for improvement in performance and cost-efficiency. Furthermore, most switching research is based on uniform incoming cell/packet traffic, which is very different from real time traffic. Real time traffic is not only bursty, but also involves multiple classes of prioritized traffic, as well as multicast traffic.

In this thesis, a high performance ATM and IP switch architecture is introduced. The switching architecture is based on two existing technologies namely Random Early Detection (RED), and the internal buffer. Simulation results show that with a little modification of these schemes, a switch can perform extremely well under many kinds of real time traffic patterns, including multi-priority and multicast. In addition, the proposed switching architecture shows that cell loss ratio can be arbitrarily reduced using a finite internal buffer size.

Table of Contents

List of Figures

List of Tables

Acknowledgment

I would like to thank my supervisor, Dr. H. M. Alnuweiri, for his constant guidance and advice in supervising my master's research work. His patience and support throughout this project are very much appreciated.

Chapter 1 Introduction

One of the most outstanding features of ATM switching is its ability to provide per-flow quality-of-service (QoS), guaranteed under a multi-rate, multi-service, variable bandwidth environment, in a scalable and manageable manner. ATM networks have the potential of subsuming the Internet and telephone network to provide a unified infrastructure for supporting guaranteed traffic (e.g. voice and video), as well as best-effort traffic (mainly data). No other technology can provide the type and versatility of QoS that ATM can provide in a scalable manner and at a high-bandwidth. Not only is ATM switching well established for the wide-area-network (WAN), it is also becoming the tool of choice for implementing the emerging technology of IP switching in the local-area-network (LAN). The underlying concept is to combine ATM as the level 2 (data-link layer) switching protocol with level 3 (network layer) IP routing in a single "IP" switch.

This new vision of the expanding role of ATM switching to provide QoS routing in LANs and WANs, requires taking a second look at what functionalities and guaranteed performance can be delivered by next-generation switches. Few vendors supply per-flow queuing chips that enable accurate QoS control at a switch port. However, these chips are planned for use in small switch configurations. For large switch configurations, many challenging problems remain to be solved, including how to route cells with minimum delay and cell loss ratio to the output ports where per-flow queuing can be exercised, and how to manage and coordinate multicasting operations in the switch fabric. The efficiency of per-flow queuing at the output ports will be greatly reduced if delay-sensitive cells are delayed considerably by the routing fabric or at the input ports. Similarly, per-flow scheduling and fair-share bandwidth allocation will be ineffective if cells experience excessive QoS coupling in the routing fabric.

1.1 Features of Next-Generation Switching Fabrics

First generation ATM switch designs were mainly focused on single QoS-class traffic. However, the integration of LAN with other broadband traffic, including voice and video, requires switches to be able to handle multi-class traffic and multicast efficiently without violating the QoS guarantees for the wide range of services. The addition of the available bit-rate (ABR) component introduces substantial complexity to the switch buffering and internal signaling structure because switches must monitor internal congestion, QoS guarantees fairness to the individual virtual channel (VC), and it cooperates with end-to-end feedback congestion control mechanisms. The following are some important qualities of next-generation switching fabric:

1. **QoS Preservation:** This requires switching fabrics to provide low cell delay, low cell loss and low cell delay-variation (jitter).
2. **Flow Isolation (or QoS Decoupling):** QoS handling is normally relegated to the buffer controllers at the output and input ports of the switch. Therefore, the core fabric should minimize the QoS interaction or coupling among flows from different input ports. Fairness is another important quality related to flow isolation. On a coarse level it means that all input and output ports should receive equal service and equal internal bandwidth. On a finer level, it implies that cells from each VC must be serviced according to some weighted fairness measure.

3.**Multicasting:** This capability is essential for next-generation switches. Multicasting is a challenging problem for large-scale switches because early replication of multicast cells can potentially congest the switch fabric.

4.**Scalability:** This is a measure of the normalized (or average) growth in hardware and buffering resources to maintain a uniform performance as the switch size or the external link rate is increased.

5.**Implementation Feasibility:** For switch architectures with multi Gb/s link rates, the feasibility of implementation is limited mainly by input and output (I/O) requirements, off-chip communication speeds, and inter-board wiring and cabling. The switch should be modular so that it can be easily partitioned among multiple chips and boards.

1.2 Motivation and Scope of Thesis

This thesis focuses on the issues pertaining to the design and implementation of next-generation large-scale ATM switches that support QoS routing and buffering, and efficient multicasting with an external link-rate reaching up to 10 Gb/s per port (i.e. OC-192 rate) or higher. At this junction, it is worth stressing that the switch core fabric and the switch ports are viewed as two distinct entities. The switch ports normally reside in the line cards and they contain the major buffering and QoS control resources of the switch, while the switch core is responsible for routing and/or multicasting cells to the output ports. Delivering high QoS performance on a per-flow basis is a challenging task especially in large-scale switches. The problem for large-scale switches is to deliver cells to the output buffers with minimum delay, and preferably, with no cell loss. Another source of difficulty is the potential occurrence of unbalanced or hot-spot traffic, such as when several workstations simultaneously burst data to a server, or when several subnets send data to an inter-switch trunk. Such traffic can easily exhaust even the most generous buffering resources at the hot-spot output port(s), and therefore, must be handled by efficient distributed buffering or by back-pressure signaling. The presence of unspecified bit rate (UBR) and ABR traffic further complicates matters. These two kinds of traffic are designed to occupy any leftover bandwidth after servicing guaranteed traffic. The presence of this "background" traffic will tend to saturate switching fabrics most of the time, especially if ABR / UBR multicasting is supported. This problem can be more serious than it may appear at first, because ABR / UBR traffic is not subjected to connection admission and control (CAC). In surveying present architectures for ATM switches, two very important qualities of switching fabrics have hardly been addressed, which are the issues of *QoS coupling* and *fairness*. QoS coupling occurs when multiple flows (e.g. traffic from different input ports) interact through the routing fabric and subsequently affect each other's performance (delay and loss). Severe QoS coupling can lead to serious performance degradation especially for delay-sensitive traffic. The issue of fairness is related to the fabric ability to provide equal service to traffic flows independent of their input or output port number.

The switch architecture proposed in this thesis is based on the approach of internal bandwidth expansion, which increases the internal bandwidth of the switching fabric either by paralleling or by speeding up the switching resources (switching elements and internal links) at each stage of the fabric. Either method implies that the output buffers must operate at a higher speed than the input link rate. The required level of speedup depends on the degree of internal bandwidth expansion. However, output buffer speedup can be controlled by using a small amount of internal buffering. Bandwidth expansion can be achieved by several methods including *fabric replication*, *fabric dilation*, or by using *tree-type* networks for cell distribution from input ports, and for cell concentration at the output ports. The proposed switch architecture is based on the fabric dilation method.

1.3 Thesis Contributions

This thesis proposes several techniques for the designing and implementation of next-generation large-scale switch fabric and the support of QoS routing and buffering. The following elaborates on several specific contributions of this thesis:

1. This thesis proposes a Buffered-Dilated Banyan (BDB) switch architecture, which provides high utilization of two main resources of the switch: interconnect and buffers. Interconnect and wiring complexity become a major bottleneck for large-scale implementations, where the switch fabric spans several boards and shelves. Thus, minimizing interconnect complexity is an important design issue. Buffer is a significant component in determining the cost and delay performance. Switch with QoS and multicast support may require large buffers in order to guarantee all the QoS requirements and multicast packet streams passing through the fabric completely. Hence, the optimization of buffering in switches is another important design issue.
2. Analyses of the performance of the proposed fabrics with computer simulations under independent and bursty traffic conditions are provided. The effects of various parameters which affect cell loss, such as internal link dilation, depth of the fabric, and share memory size, are considered in the performance analysis.
3. Different buffering strategies on the proposed architectures are studied. This is based on the use of a truly self-routing interconnect. Input, output and input-output buffering are considered. The notion of the link-capacity back-pressure mode of operation for all different buffering strategies is also introduced. This

mechanism eliminates cell loss from the internal stages of the switch and limits cells to the BDB switch inputs only. In the switch core, a small amount of strategically placed buffers can be used to control cell loss and enhance switch robustness to changing traffic characteristics.

4. Multicast traffic is a significant portion of all of the existing traffic. A multicast mechanism is introduced into the switch as a copy network. Its performance under multicast traffic on both the uniform and bursty traffic pattern is analyzed by simulation. The proposed copy network achieves very low cell loss even at very high output offered loads.

1.4 Organization of the Thesis

Chapter 2 presents a comprehensive survey of the existing switch architectures. It takes an in-depth look into the existing switch fabrics and highlights their drawbacks and the required improvements for large-scale switching.

Chapter 3 looks at the architecture of the Buffered-Dilated Banyan switch in terms of the interconnection complexity. It also addresses the features of the BDB switch.

Chapter 4 looks at the buffer management of the switches. The ideas of Random Early Detection (RED) and Drop Tail Mechanism are introduced. The overall system performance is analyzed by computer simulations.

Chapter 5 introduces the input-output buffered BDB banyan. The performance under uniform traffic conditions is analyzed. The performance of the input-output buffered banyan is further enhanced by placing a small amount of internal buffers into the switch element and introducing a backpressure mechanism. Simulation results under independent uniform and bursty traffic conditions with multicast cells are presented.

Chapter 6 concludes the thesis and outlines future research areas.

Chapter 2 Switch Architecture

2.1 Switch Element

Switching elements are building blocks of switching fabric which is the core of an ATM switch. In this section, four ATM switching elements will be presented. They are Knockout, Roxanne, Coprin and Arena [7] [22]. Although they do not represent all of the ATM switching elements, their architectures are very different.

2.1.1 Knockout Switching Element

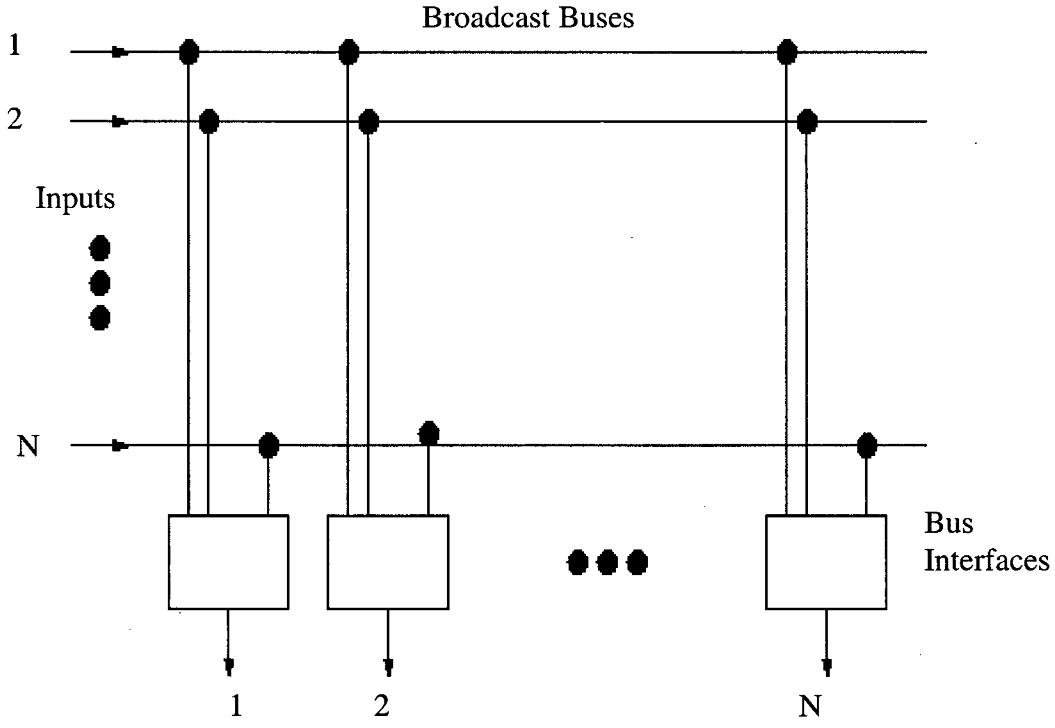


Figure 1 : Knockout Switching Element

The Knockout Switch belongs to the crossbar network family used to describe any single path non-blocking network that has a complexity, and which grows as a function of N^2. Figure 1 shows a simple block diagram of the Knockout Switch with N inlets and N outlets (NxN). One of the advantages of the Knockout Switch architecture is that each bus is driven by only one inlet instead of sharing multiple inlets. This makes implementation simpler, and makes a higher transmission rate possible.

Since N inputs are connected to one outlet via a bus interface, N cells may contend for the same output. Therefore, the bus interface provides a queue. This is the reason why the Knockout Switch is based on an output queuing solution. Figure 2 shows a closer look at the bus interface, which contains the cell filters, the concentrator, and the shifter.

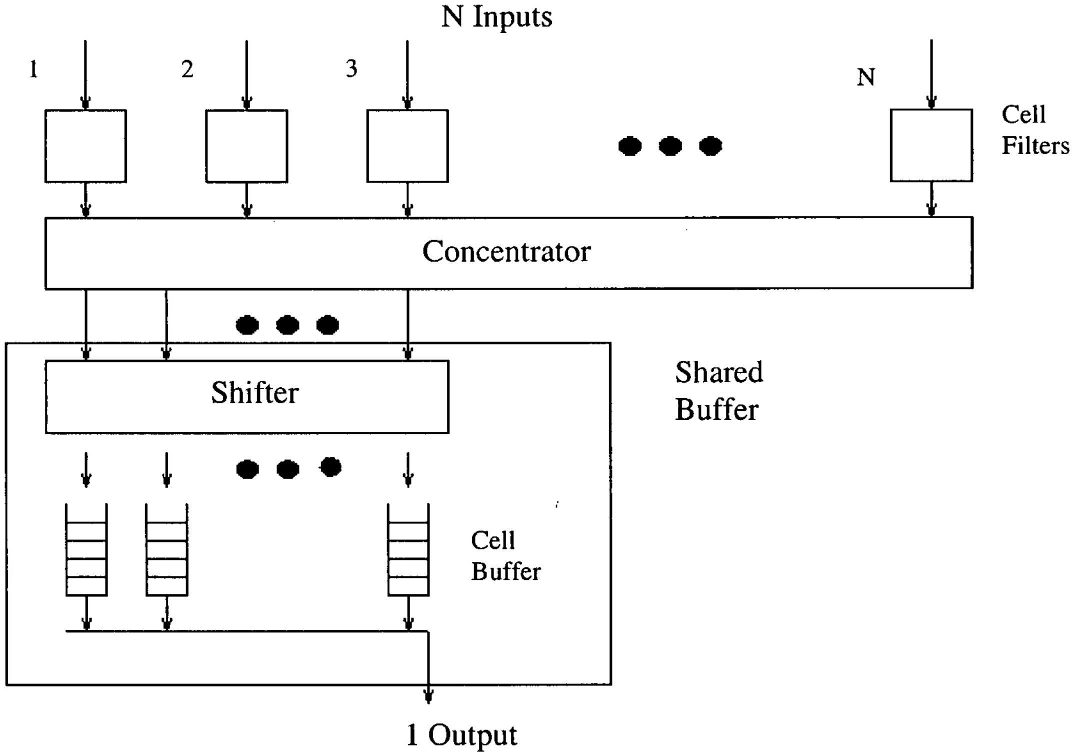


Figure 2 : Knockout Bus Interface

Due to the connections on the bus interface, N cells may arrive at one bus interface and try to reach one particular output. Without the cell filters, concentrator and shifter, the buffer must operate at N times the speed of one inlet in order to send out all the cells at one cell time. However, this may put too much constraint on the memory. Instead, the Knockout Switch uses a smarter bus interface, which lowers the required memory. By scanning the addresses of incoming cells, the cell filters select ones that are destined for specific outlets and let them pass, otherwise the cell is discarded. Then, the concentrator concentrates N inputs to L outputs. If the number of incoming cells is greater than L, some of the cells are dropped; however, the probability of this is very small.

The purpose of the concentrator is to reduce the number of memory writes to the buffer in one cell time to L instead of N. In addition, the concentrator reduces the number of buffers required. The concentrator is divided into several stages composed of many contention switches. Figure 3 shows one block of a 2x2-contention switch, and Figure 4 shows the architecture of a concentrator.

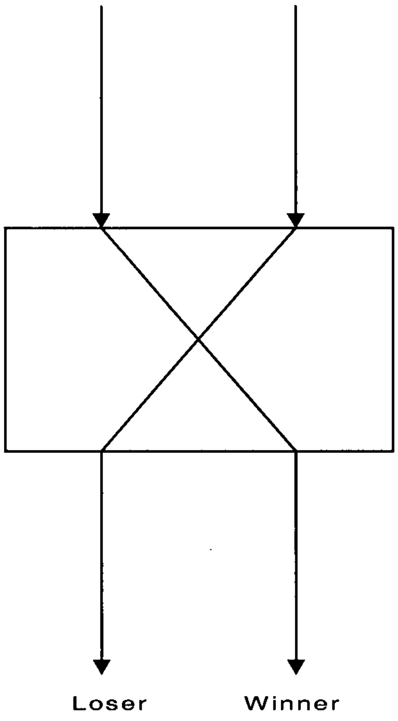


Figure 3 : Contention Switch

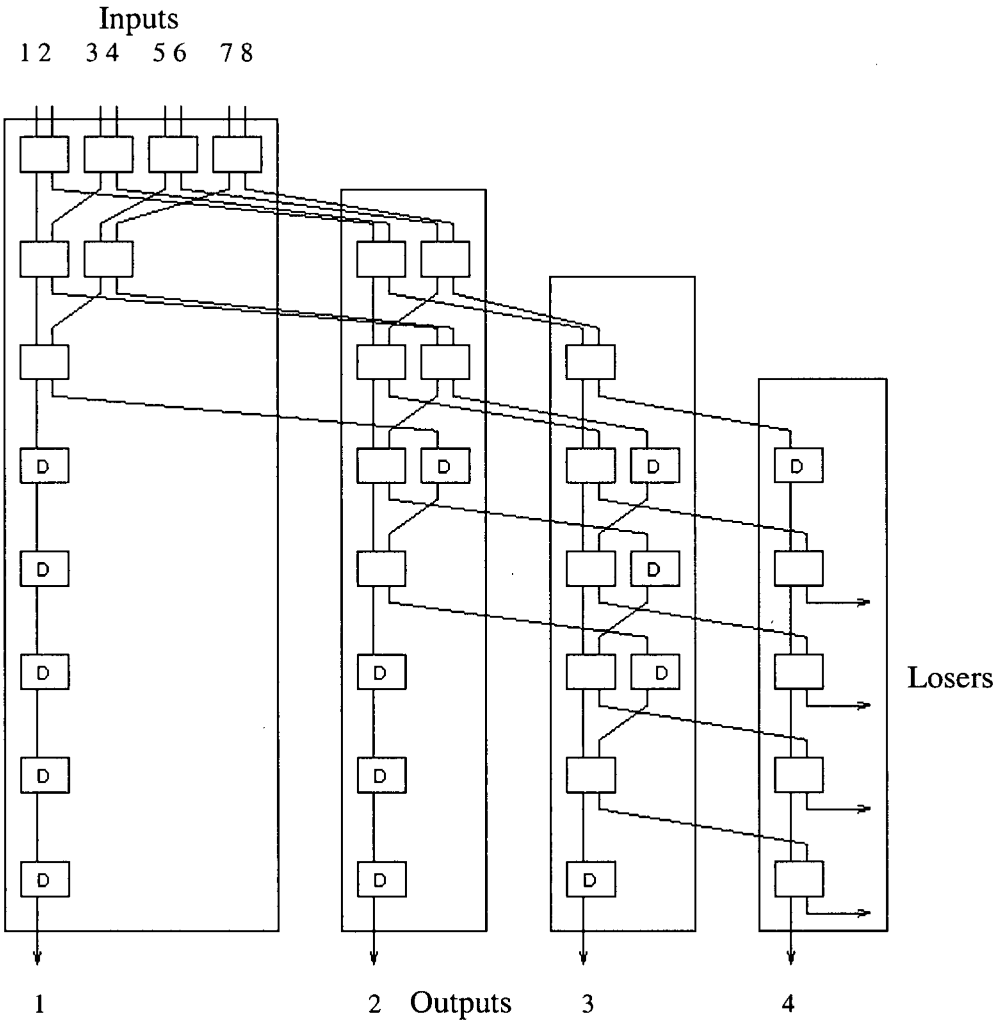


Figure 4 : An 8 input/4 Output Concentrator

The contention switch operates in the following ways:

1. If only one cell is present in one of the inputs, it is selected as the winner (pass).
2. If two cells are present in both of the inputs, then the left cell is selected as the winner and the right cell is dropped.

In order to implement an 8-input, 4-output concentrator, the 2x2 contention switches are organized in a Knockout game architecture. The Knockout game can be broken down into different stages (a tournament). In each tournament, players (incoming cells) are knocked out of the tournament as soon as they lose one match, and one winner is selected. The losers in the first tournament again compete in the second tournament, and the winner becomes the second in rank. Finally, in the last tournament, the losers are discarded.

When there is an odd number of players in one round, the odd player must wait for a competitor. This requirement can be achieved by adding a 1 bit delay element indicated by a letter D as shown in Figure 4.

An interesting property of the concentrator is that if the number of input cells is less than L, the concentrator always concentrates the cells in the left most outputs.

The cell buffers are separated into L queues. This allows for a simpler implementation of the buffer because now each queue only needs to handle one write and one read in one cell time, instead of L writes and one read in a combined queue implementation. Less memory access means that the memory does not require fast access time.

The shifter in front of the cell buffer guarantees that all the L buffers are equally loaded and optimally used. Moreover, the cell sequence at the output is guaranteed in this way. The shifter is a circular shift register. Figure 5 shows two instances of its operation.

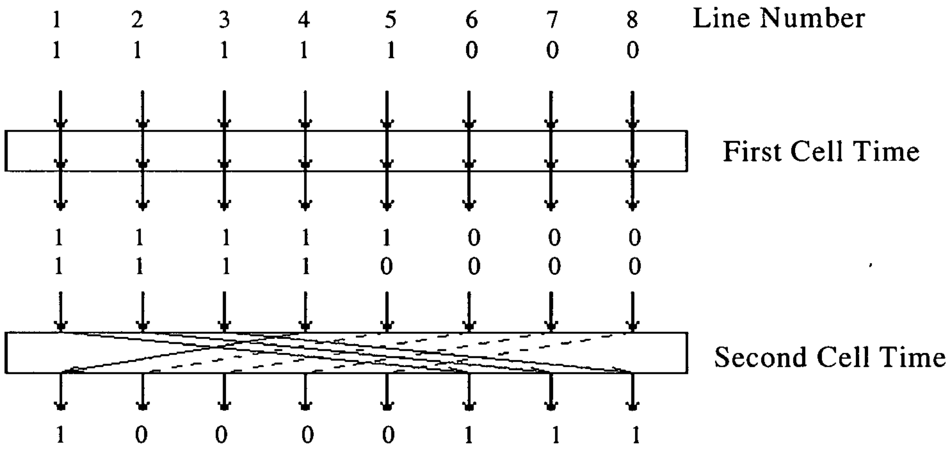


Figure 5 : Shifter Function for 8 Lines

In this figure, “1” indicates that there are cells on the inputs. In the first cell time, input one is mapped to output one, and so on. In the second cell time, input one is mapped to output 6, since output 5 was the last output filled the time before. In a more mathematical notation, if S_i denotes the position that the shifter must shift to on the right during cell time i, then the following is true:

$S_{i+1} = (S_i + k_i) \bmod L$

where k_i represents the number of filled cells arriving during cell time i. It is assumed that $S_1 = 0$. For example, using the situation shown in the figure, let $k_0 = 5$. Thus, the position of the shifter at the next cell time (1) is $S_1 = (0 + 5) \bmod 8 = 5$. Therefore, the shifter is now at line 6 (position 5, starting from 0).

Multicast/broadcast Capability

Since each bus interface has already connected to N inlets, and the multicast/broadcast information gets distributed over all of the bus interface due to this configuration, it is quite easy for the Knockout Switch to perform a multicast or broadcast function. The only requirement is that the cell filters must be able to distinguish the multicast/broadcast cell and know if the incoming cell belongs to its outlet. In order to avoid this complex functionality, which is built on all of the cell filters in every bus interface, multicast modules are added to the bus interface network, as shown in Figure 6 [22].

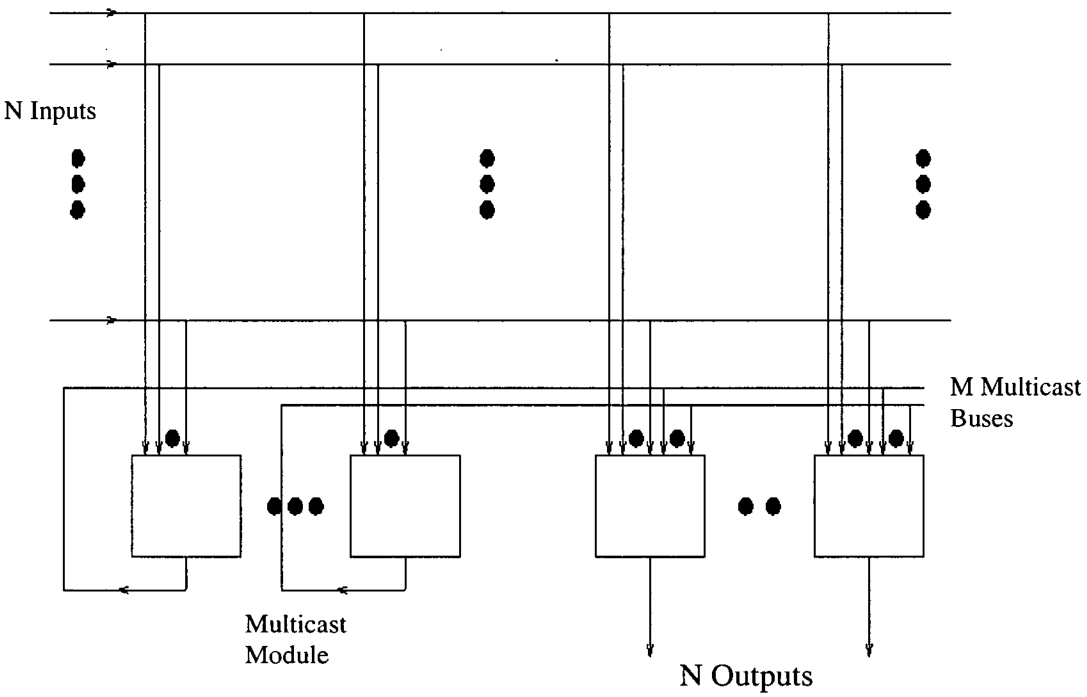


Figure 6 : Knockout Switch with a Multicase Module

The multicast modules have very similar structures to the bus interfaces. Moreover, the multicast module has a cell duplicator and a table of multicast virtual circuits. The architecture of a multicast module is shown in Figure 7.

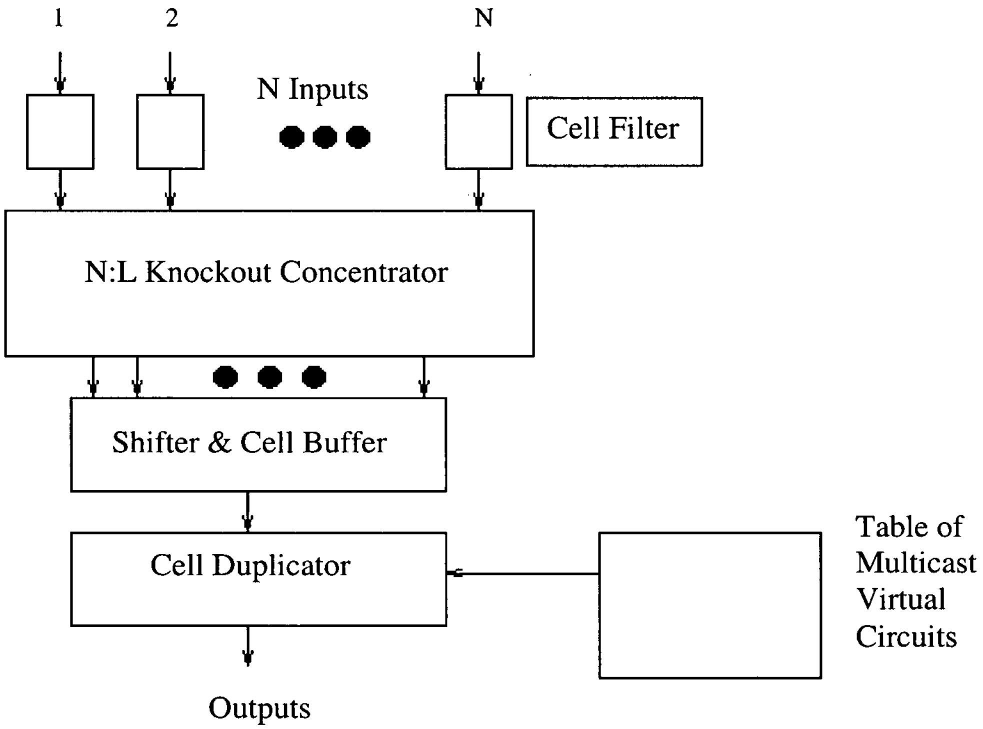


Figure 7 : Multicast Module with a Cell Duplicator

In the multicast module, the cell filters, concentrator, shifter, and cell buffer are all the same as the one in the bus interface described previously. The table of multicast virtual circuits provides the number of copies required for a particular multicast/broadcast cell and the destination address of each copy. The information in the table is updated at connection time. Having received that information, the cell duplicator makes the correct number of copies and sends them out to the bus interface via the multicast buses. If there are M multicast modules, the bus interface requires N + M inputs. The number M depends on the demand of multicast functionality.

2.1.2 Roxanne Switching Element

Invented by Alcatel researchers in 1990, the design of the Roxanne Switching Element is based on the central queuing principle [22]. The switching element is called the Integrated Switching Element (ISE), and it supports a multicast/broadcast function. One possible dimension of the ISE is a 32 x 32 basic switching block with inlets and outlets operating at 150 Mbit/s. Figure 8 is a block diagram of an ISE.

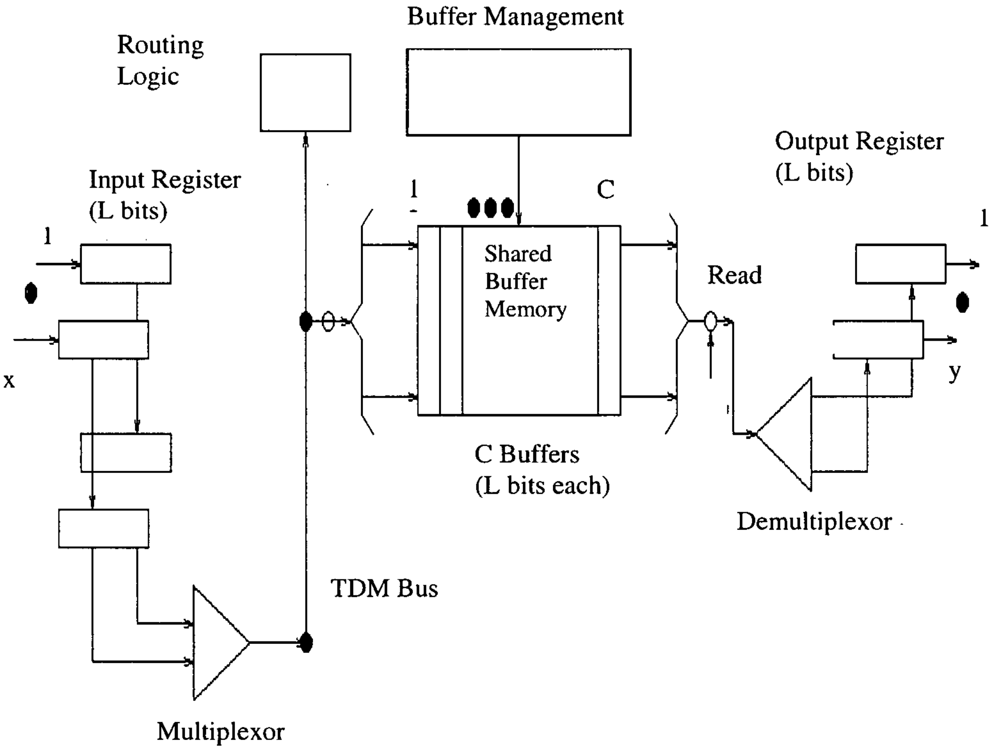


Figure 8 : ISE Function Block Diagram

After L bits of data have been shifted into the input register (serially), the L bit data is latched. Then the data (parallel) is sent to the Shared Buffer Memory (SBM) via the TDM bus (L bit wide). The conversion from serial input to parallel input lowers the requirement for the internal operating speed. If the inlets and outlets transfer at 150Mbit/s and the TDM bus is 8 byte wide, the frame period is 64bit/150 Mbit/s = 0.43μs. If the ISE has 32 inlets, 32 write operations have to be performed by the SBM within one frame period, which is the time taken to fill up the input register. Therefore, the internal bus speed becomes1/0.43μs x 32 = 74.4Mwords/s (word size is 8 bytes).

In order to use the memory space more efficiently and reduce memory size, queues in the SBM represent groups of 4, 8, 16 or 32 outlets. A queue such as this that is called a "logical queue" and is treated like a single queue. The idea is to distribute the load as equally as possible on all outlets. Before a cell is written into the ISE, the routing logic performs tag processing which mainly consist of routing information interpretation. Implemented as a pipeline machine, it analyzes the self-routing tag on the cell and routes the cell to a logical queue, depending on the specified ISE routing mode. Possible routing modes include direct routing modes, where a particular outlet is specified, and distributed routing mode where a cell may be routed freely to a group of different outlets. The use of a distributed routing mode illustrated in the Roxanne Switching fabric section.

The buffer management keeps track of the free and occupied cells by using a linked list. In addition, it is responsible for handling the multicast/broadcast function.

Multicast/Broadcast

For multicast cells, all routing tags contain an internal reference number. The reference number is used to access a special memory (not shown in Figure 8) in the ISE, where the number of copies and the destinations of the copies are stored. The destinations are represented as a mask. This mask comprises one bit per logical queue, indicating whether or not a copy has to be placed in that logical queue. The buffer management allocates one copy count per cell. When a copy is sent, the count decreases until the count is zero. Then, the address of the cell can be released.

Multislot Cells

An interesting point about the Roxanne Switch is that it will convert incoming cells into several fixed length slots before they are stored. Within each slot, 2 bits are used to indicate the slot sequence (start of cell, end of cell). As a result of this data structure, the Roxanne Switch is compatible with variable cell lengths. This may extend the application of the Roxanne Switch to other types of networks. Another advantage of this structure is that a new cell can start to move into the slot of a previous cell location even though not all of the slots of the old cell have become available. Consequently, memory space can be used more efficiently. However, if there are n slots in a cell, the size of the linked list management in the buffer management will be n times larger.

2.1.3 Coprin Switching Element

The Coprin Switch was discovered and designed by the French CNET in order to transfer data, voice, and video signals. It was originally intended to operate with links at 280 Mbit/s and cells of 15-byte information, and a 1-byte header that contains routing information. It has a reference number that is known internally to the switch and is determined at the time of connection. The Coprin Switch uses the reference number to determine the physical outlets of a cell. The design of the Coprin Switch is based on the central queuing system. The block diagram of the Coprin switch is show in Figure 9.

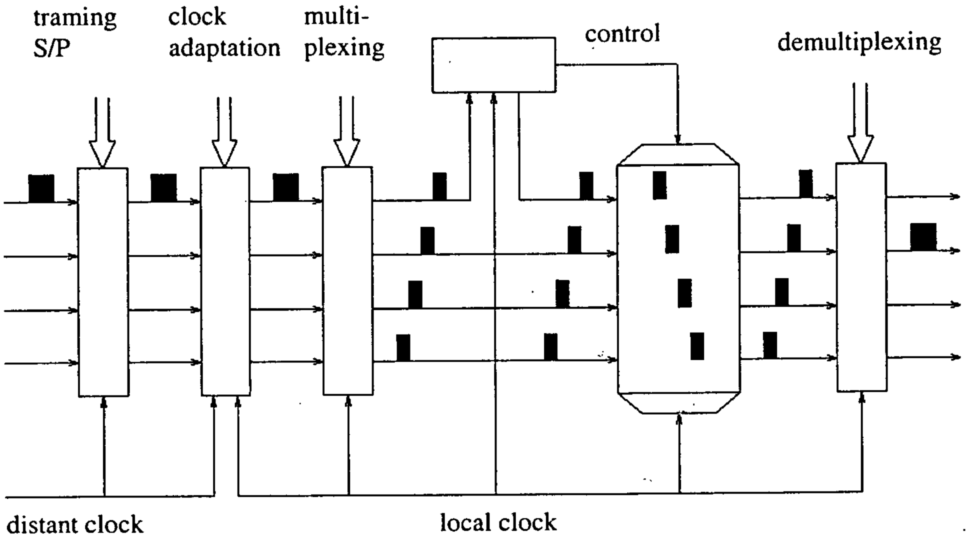


Figure 9 : Coprin Switch

A major characteristic of the switch is that a cell coming into the Coprin Switch will be converted to a parallel stream of information inside the switch. The Super Multiplexing block is responsible for this task. Figure 10 shows an example of a 4 by 4 super multiplexing function, assuming that each cell has one header byte and 3 data bytes.

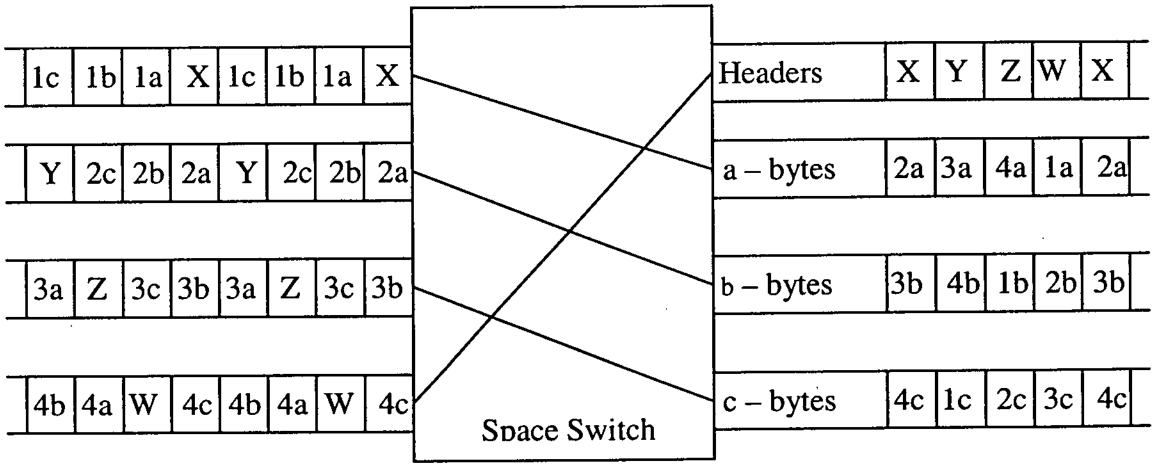


Figure 10 : Extracting Packets in Buffer Memory

In order for the Super-Multiplexing block to function correctly, the headers on different inlets have to arrive in consecutive time slots, as shown in Figure 10. This alignment of the incoming cells is an important precondition. In order to satisfy this requirement, the phase alignment block is put before the Super-Multiplexing block, as shown in Figure 10.

Super-Multiplexing

The function of the Super Multiplexing Switch is to place bytes of information of different meaning into different channels. Through this way, different parts from different cells are loaded into the memory at the same time. Therefor, the cell loading operation is "pipelined". As shown in the example, headers are put on the first outlet, the second byte is put on the second outlet in the next clock cycle, and so on. Therefore, consecutive slots in the outlets actually contain information from different inlets. In order to achieve this function, the Space Switch is implemented in four states as shown in Figure 11.

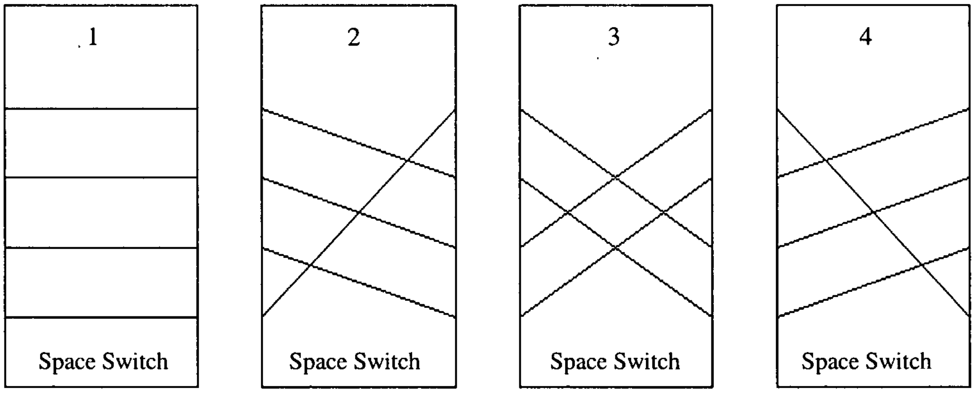


Figure 11 : Four States of the Space Switch

The four states will be rotated on each clock signal (in the order 1, 2, 3, 4, 1...).

Buffer Memory

The buffer memory stores the ATM cells in parallel form. The central queue is divided into banks that are dedicated to a byte of the cell. For example, if the header address of a cell is A and it is stored in bank 1, then the first data byte will be located at A +1 which is in bank 2. Figure 12 below, shows how packets are extracted in the buffer memory.

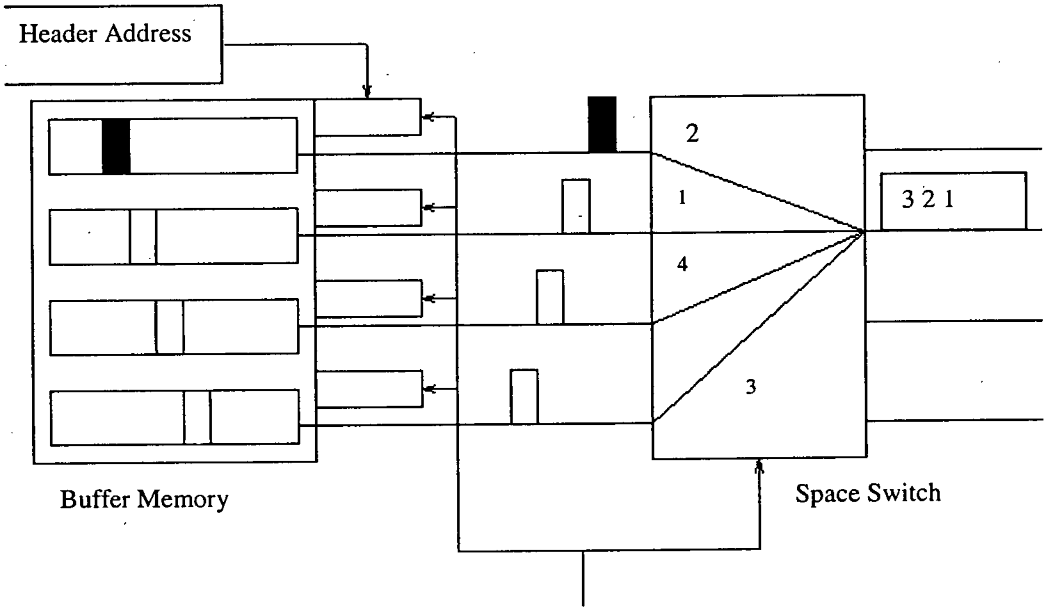


Figure 12 : Extracing Packets in Buffer Memory

First the header address (A) will be provided by the control and sent to bank 1. Thus, the header stored in A will be retrieved. In the next cycle, the address A + 1 will be sent to bank 2 and the data in address (A + 1) will be retrieved. The header address will continue to ripple down into the next bank with the correct index on each clock cycle until it has reached bank 4.

Before the data is put on the outlets, the ATM cells are demultiplexed. In order to reconstruct the cells, the demultiplexor performs the opposite task of the Super-Multiplexing module. It has a Space Switch of the four states described in the Super Multiplexing. However, for the demultiplexor, the four states are rotated in the opposite direction, that is, 4, 3, 2, 1, 4 and so forth.

Routing operation

The routing operation is done in the control block. It consists of a translation table, where each item has 3 bytes of data, 1 byte of translated header, and 2 bytes of destination locations. When a cell arrives, the control will determine the translated header of the cell and replace the cell's incoming reference number with the translated header. This is accomplished according to the time basis (which tell from which inlets the cell arrives) and the value of the incoming header. The new header will write into memory to where the header should be stored. The other two bytes of data in the table are used to indicate whether a cell has to be copied to any of the 16 outlets by allocating a '1' to the specific location represented by an outlet. Figure 13 shows the basic structure of the control.

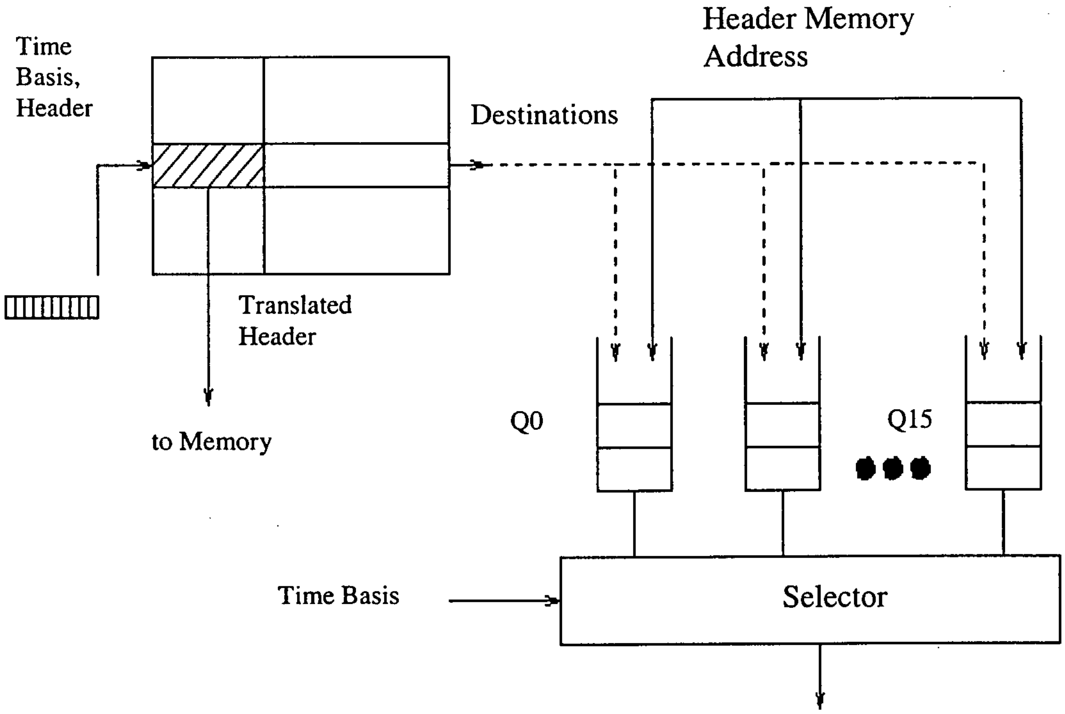


Figure 13 : Control of the Coprin Switch

The header address in the memory is stored in the queues (Q$_0$ to Q$_{15}$) if the cell has to be copied to those specific outlets. The advantage of using the queues to hold the header addresses is that the output contention problem can be avoided.

An Example of a Routing Operation

Please refer to Figure 13 as an example of the arriving cell patterns.

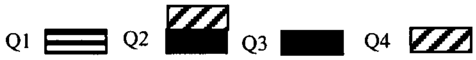
Suppose those headers have the following destinations and addresses in the memory:

■ 3, 2 :Addr 10

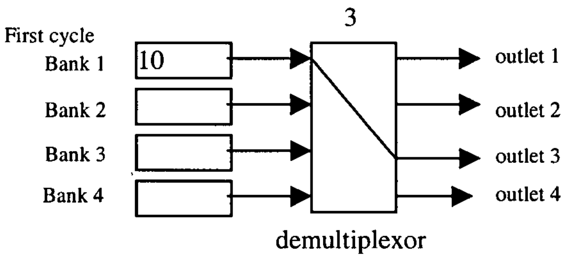
▨ 2,4 :Addr 20

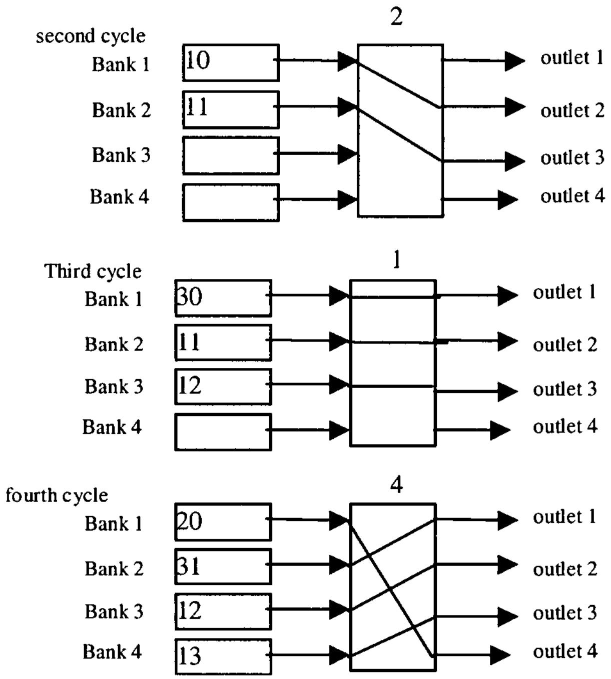
▤ 1 : Addr 30

After all of the three headers have been scanned by the control, the queues will look like this:



Below shows the addresses on each bank in the memory after the headers are ready to be retrieved from the memory. The selector shown in Figure 13 will select the header address value in the queue whose number is the state number in the demultiplexor. For instance, if the demultiplexor is in state 3, the selector will pick the address in queue 3. Suppose the selector starts with queue 3, then after the selector has picked up the header address, the address will be removed from the queue.





As shown in the example, the broadcasting or multicasting operation is quite simple in the Coprin Switch.

2.1.4 Athena Switching Element

The Athena Switching Element is based on the output queuing principle. The switch was originally designed on a single circuit board with 16 inlets and 16 outlets which are operate at 600Mbit/s. Since the transfer medium is fully non-blocking, it must be capable of operating at 9.6 Gbit/s (16 x 600Mbit/s). The high throughput can be overcome by adopting cell slice architecture. Figure 14 shows the internal architecture of the Athena switch [22].

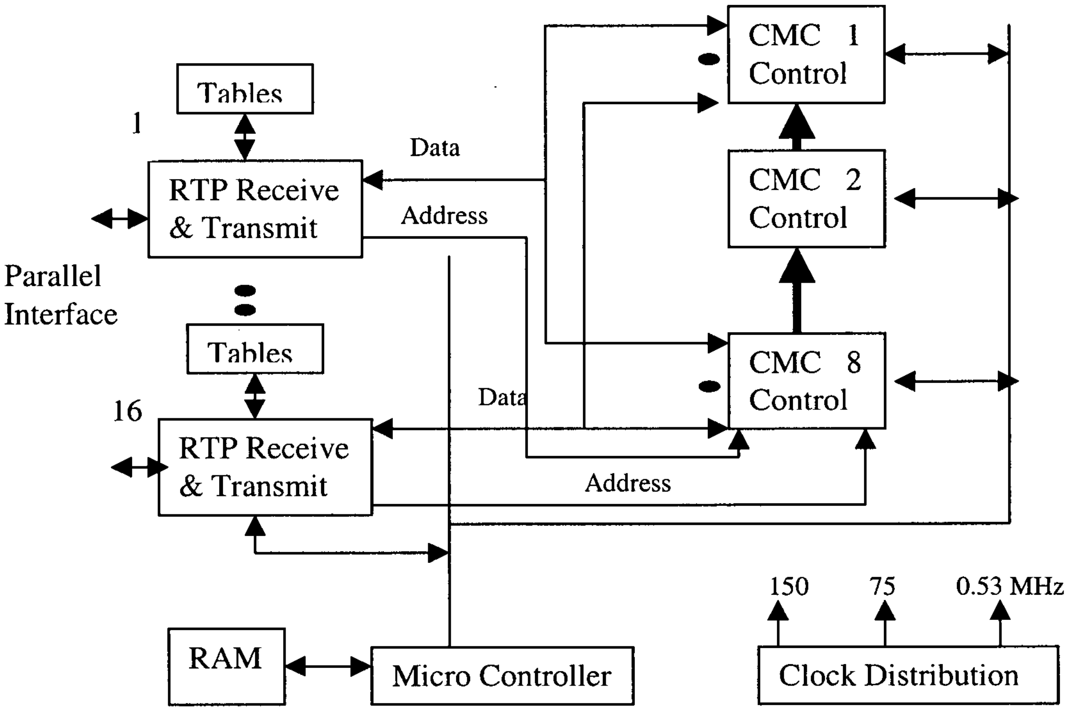


Figure 14 : The Athena Basic Switching Block

Each parallel interface is an input port. The output queues are broken down into 8 chip called the Central Memory Chip. Each chip handles one eighth of cell information and is connected to 16 Receive and Transmit ports (RTP). Connected to an inlet, each RTP performs cell reception, header processing (including header error detection /corrector), label translation, maintenance, and cell transmission to the next stage of the switching fabric. The RPT will determine the designated outlets of the cells. Since each cell is distributed to 8 CMCs, the internal bus speed can be reduced to 75Mbit/s (600Mbits/s divided by 8).

The micro-controller is connected to all RTPs and CMCs via a special data bus in order to update the routing tables in all of the switching elements. The routing information will be transmitted internally in the switching fabric by control cells. The control cell is characterized by a special header value. In addition, the micro-controller can interpret the priority bit of a cell.

Central Memory Chip

The architecture of a CMC is shown in Figure 15 [22].

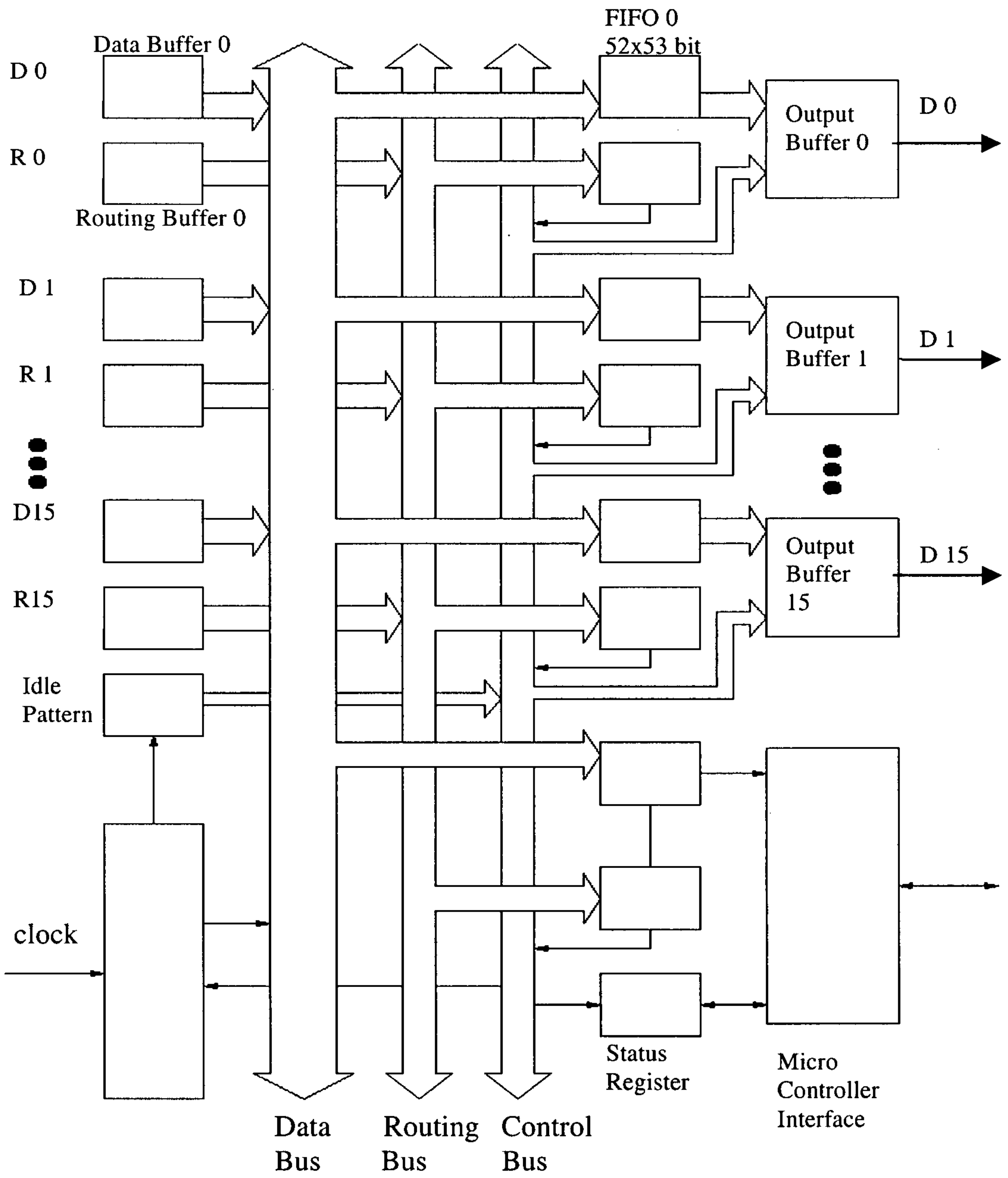


Figure 15 : The CMC of the Anthena Switch

D1x and Dx are actually the same bi-directional data buses. Rx represents the address routing bus from RTPx. Each CMC has 16 data input (DI0 to DI15) and 16 address routing input buses (R0 to R15). Both the data input and the address routing input are buffered at the input. Since each CMC handles 1/8 of data from the input, 53 bits of data will be transferred to the FIFO each time. The FIFO has a pointer to the first and the last element. Each FIFO has 52 locations, 47 cell locations and 5 locations for high priority cells. The address routing inputs have 18 bits: 16 bits to represent the destination where each bit identifying whether a copy to a specific outlet has to be made, 1 bit for priority and 1 bit for control cells. The data in buffer x will be delivered to output Dx according to the routing information in the address routing buffer. An extra FIFO (FIFO 16) is used to transfer the control cell to the micro-controller on board. The sequence diagram below shows the flow of cells.

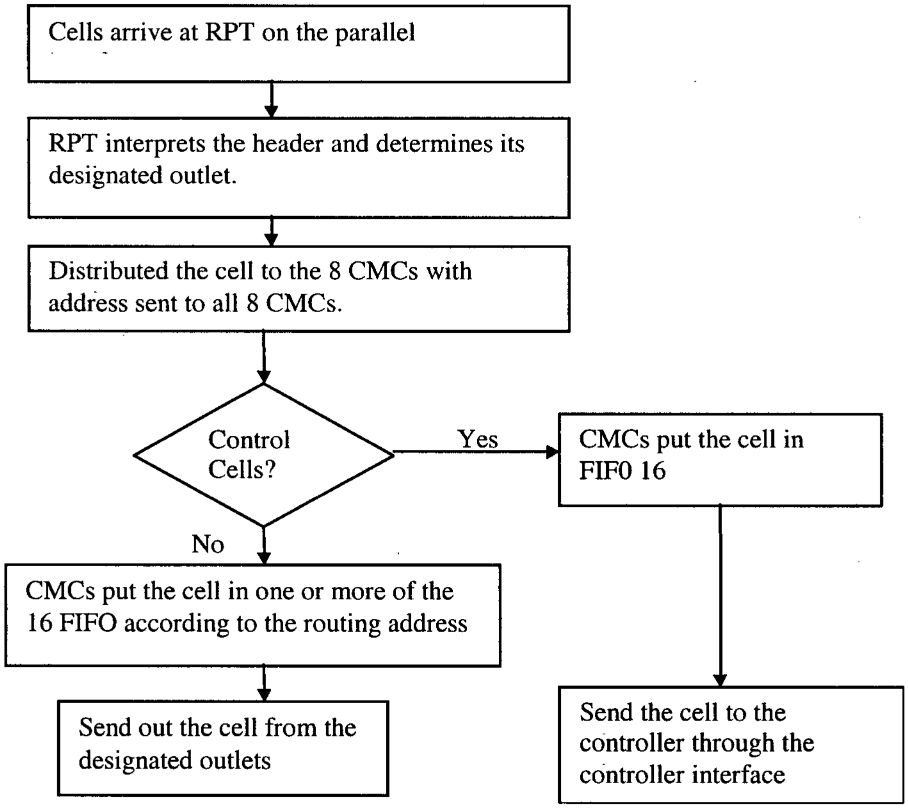

Cells arrive at RPT on the parallel
RPT interprets the header and determines its designated outlet.
Distributed the cell to the 8 CMCs with address sent to all 8 CMCs.
Control Cells?
Yes
CMCs put the cell in FIF0 16
No
CMCs put the cell in one or more of the 16 FIFO according to the routing address
Send out the cell from the designated outlets
Send the cell to the controller through the controller interface

Figure 16 : Sequency Diagram Showing how Athena Switch Handles Incoming Cells

The cell loss rate for normal cells and priority cells at 80% load is about 10^{-10} and 10^{-15}, respectively. The most important shortfall of the Athena architecture is its high number of interconnection wires. However, this is the price to pay for a lower internal speed.

2.1.5 Summary of Different Switching Elements

The following table summarizes the different switching elements in terms of cost and performance.

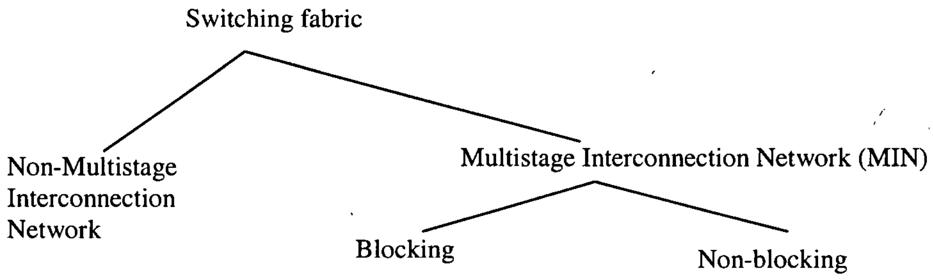
	Pros	Cons
Knockout Switching element	• Each inlet has only one bus: This design allows for simpler implementation since there is no need to consider the timing between inputs like the TDM switch. In addition, it allows for a high transmission rate • Easy to perform multicast or broadcast functions • Simple architecture	• Internal bus speed must be equal to the input/output ports' speed
Roxanne Switching Element	• Lower internal bus speed due to parallel data manipulation inside the switch • Memory space is used more efficiently by grouping queues into logical queues • Allow variable cell length by using multislot cells. • Better memory space usage and less waiting time for new incoming cells with multislot cells	• More complex memory management • May need a faster memory since the memory required to perform N reads and N writes in one cell time

Coprin Switching Element	- Simple memory management (using only a circular buffer) - Simple technique for broadcasting and multicasting - Simple design so that high multiplexing and switching capabilities can be achieved	- Requires extra circuit to perform phase alignment on incoming cells - Cell sizes are fixed - The size and number of packets in a cell is fixed since the header size defines the packet size.
Athena switching Element	- High transfer rates at output but require low internal speed - High throughput due to the highly parallel memory structure - Simple memory control mechanism since it is based on the output queuing principle	- Large number of interconnections - Large area used in memory buffer - Inefficient use of memory space since this approach is based on the output queuing method

Table 1: Summary of Different Switching Element

2.2 Switch Fabric

Switching fabric is an interconnected network composed of identical basic switching building blocks. Usually a switching fabric has a much larger number of inlets and outlets than the original switching element. Switching fabric can be categorized as the following:


Switching fabric
Non-Multistage Interconnection Network
Multistage Interconnection Network (MIN)
Blocking
Non-blocking

2.2.1 Non-Multistage Interconnection Networks (MIN)

The Knockout Switch is an example of non-MIN. An example of the Knockout switching fabric is shown in Figure 17.

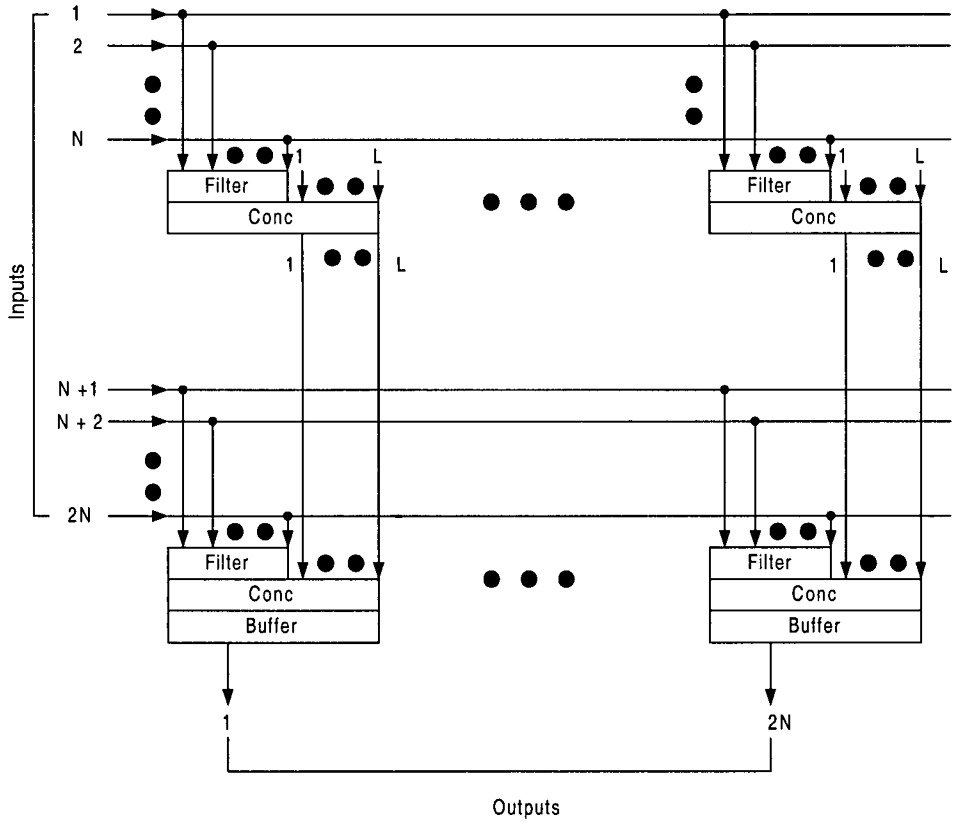


Figure 17 : 2N x 2N Switch Built with N x N Knockout Switches

This 2N x 2N Knockout switching fabric is built with an N x N Knockout Switching Element. The concentrator input now expands to N + L. For the top half of the concentrator, only N inputs are used. For the lower half, all of the N + L inputs of the concentrators in the lower half are used. The basic architecture remains the same as the single Knockout Switching Element. The shortfall of this architecture is that the size of the circuit will be increased in the order of N^2 as the number of inputs increase.

2.2.2 Multistage Interconnection Networks (MIN)

A Multistage Interconnection Network is a switching fabric that is composed of a large number of identical basic switching building blocks. Introduced by Goke and Lipovski in 1972, the Banyan network is a famous MIN [22]. In a Banyan network, there exists exactly one path from any input to any output.

2.2.2.1 Delta Networks

Making use of the Banyan network properties, Delta Networks have a self-routing property. As shown in Figure 18, no matter where the cells enter the network, they reache outlet 1011 based on the cell's routing tag. The 2 x 2 switch will look at the first bit in the tag and route the cell in the lower outlet if the value is '1', otherwise, it will route the cell to the upper outlet. In addition, after the cell has passed the first switch, the second bit in the tag will be shifted to the front.

In general, if the switching element has the dimension b x b, it requires $\log_b N$ stages, and each stage has N/b elements.

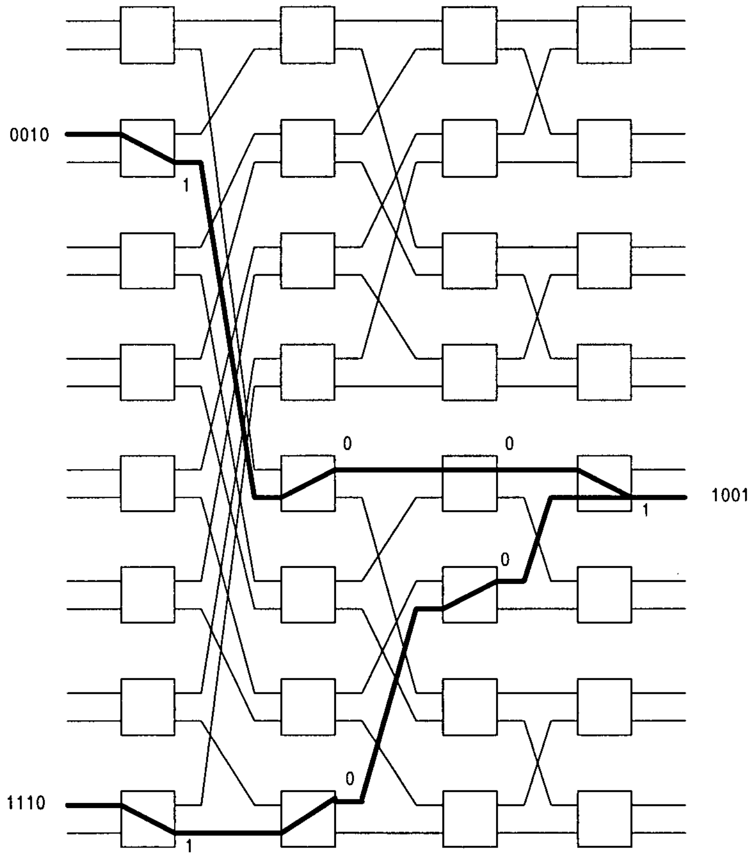


Figure 18 : Self-Routing Properties of a Delta Network

The self-routing property makes architecture simple and is suitable for high speed switching.

On the other hand, contention problems can occur at the outputs and even internally, as shown in Figure 19. Internal contention can result in internal blocking.

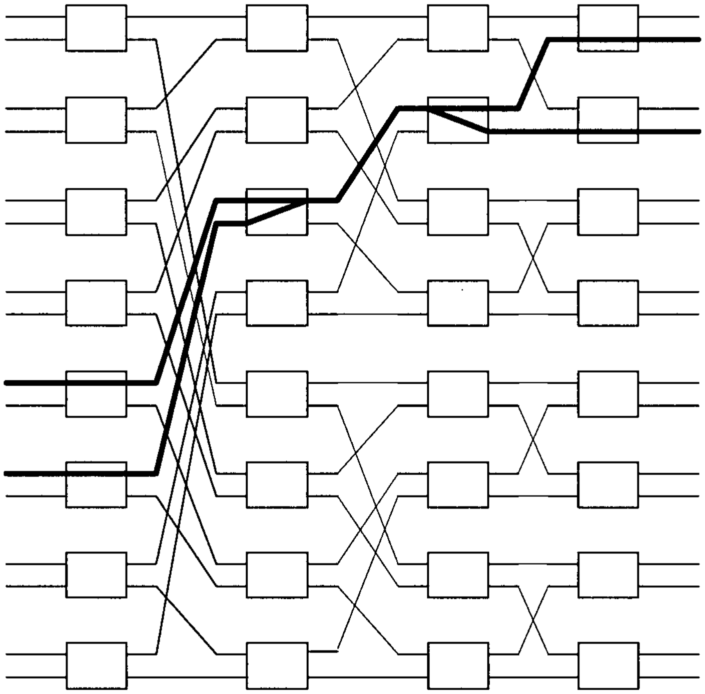


Figure 19 : Contention in a Delta Network

Therefore, the Delta Network can not be directly used as an ATM switching fabric. One must solve the internal blocking problem first. The following are some ways to reduce internal blocking:

1) Implement a buffer in the switching element.
2) Increase the internal link speed relative to the external one.
3) Delay the transfer of the blocked cell by using a backpressure mechanism between nodes.
4) Use a multiple plane network
5) Provide multiple links between inlets and outlets. This may lead to an out-of-sequence arrival at the output.

Most switching fabrics reduce the problem of internal contention based on one or more of the methods described above. In the next sections, MIN with internal blocking and non-blocking will be presented

2.2.2.2 MIN with Internal Blocking

MIN with internal blocking can be categorized into four types based on where the routing information is stored and when the routing decision is made, as shown in Table 2 below.

Routing descision time	**Routing Information place**	
	Cell based (Routing tag)	**Network based (Routing tag)**
Connection based	**I**	**III**
Cell based	**II**	**IV**

Table 2 - Four Types of MIN with Internal Blocking

Most switching fabrics will use only one type of solution. However, it is also possible to mix the different types of techniques, for example, some switching fabrics use one type for unicast and the other type for multicast/broadcast in order to combine the advantages of the two.

2.2.2.2.1 Comparison of Different Types of MIN with Internal Blocking

Type	Pros	Cons
I	• Cells arrive at the output in sequence since each connection has a fixed path in the MIN • No routing table is required	• Routing tags may increase the overhead of transporting cells. More overheads are required in this type than type II since type I needs routing tags that contain routing information of all stages. • Multicasting is more difficult to perform.
II	• No routing table is required • More efficient use of resources in the MIN since resources are shared between all links • Internal traffic characteristics are independent of the external traffic characteristic due to the randomization process.	• Cells may arrive out of sequence since cells may have a different path through the network for the same connections. • Routing tags may increase the overhead of transporting cells. Less overhead is required in this type than in type II since some stage are used for randomization. • Multicasting is more difficult to perform.

III	• Cells arrive at the output in sequence since each connection has a fixed path in the MIN • Easier to implement the multicast function by using routing tables.	• Requires routing tables
IV	• More efficient use of resources in the MIN since resources are shared between all links. • Internal traffic characteristics are independent of the external traffic characteristics • Easier to implement the multicast function by using routing tables.	• Cells may arrive out of sequence since cells may have a different path through the network for the same connection. • Require routing tables

2.2.2.2.2 Roxanne Switching Fabric

The Roxanne switching fabric uses type II routing for point-to-point connections and type IV routing for multicast/broadcast. The capacity of each link is 150Mbit/s while the maximum number of links is 16K. Although it is one of the MIN with internal blocking, the cell loss rate of a Roxanne Switch module of 128 x 128 is lower than 10^{-11}.

Since the network is type II, the routing decision is maded cell by cell and the routing information is stored in the cell. Therefore, there is no need for an internal connection set-up and the allocation of internal resources. As a result, the connection set-up time will be very small and the resources will be optimally shared.

The Roxanne Switch has the following features:

1) Internal queuing is used in the basic switching blocks
2) Multiple planes (networks) are used to ensure high reliability and optimal use of the switching elements.
3) Multiple paths are provided in the network so that traffic will be distributed evenly on each plane. In addition, the Roxanne switching fabric is also called the "multiple path self routing switch" since the routing information is stored in the cell.

The multiple paths and multiple planes solutions make the cell arrival rate at the basic switching block memoryless, that is, fully geometrically distributed. Consequently, the memory requirement in each switch element will be minimal. On the other hand, since cells may arrive out-of-sequence via different paths, a resequencing function will be needed at the edge of the switch. This issue will be discussed in a later section.

The configuration of the Roxanne Switch can vary with links load and number of links. Table 3 shows this relationship.

Type of extension	Expandability
Capacity(150 Mbit/s links)	16K links max. with 3 stages(each plane) 1024 links max. with 2 stages(each plane) 256 links max. with 1 stages (no plane)
External link load	0.8 max. with 16 planes 0.6 max. with 12 planes 0.4 max. with 8 planes

Table 3 - Roxanne Switch Expandability

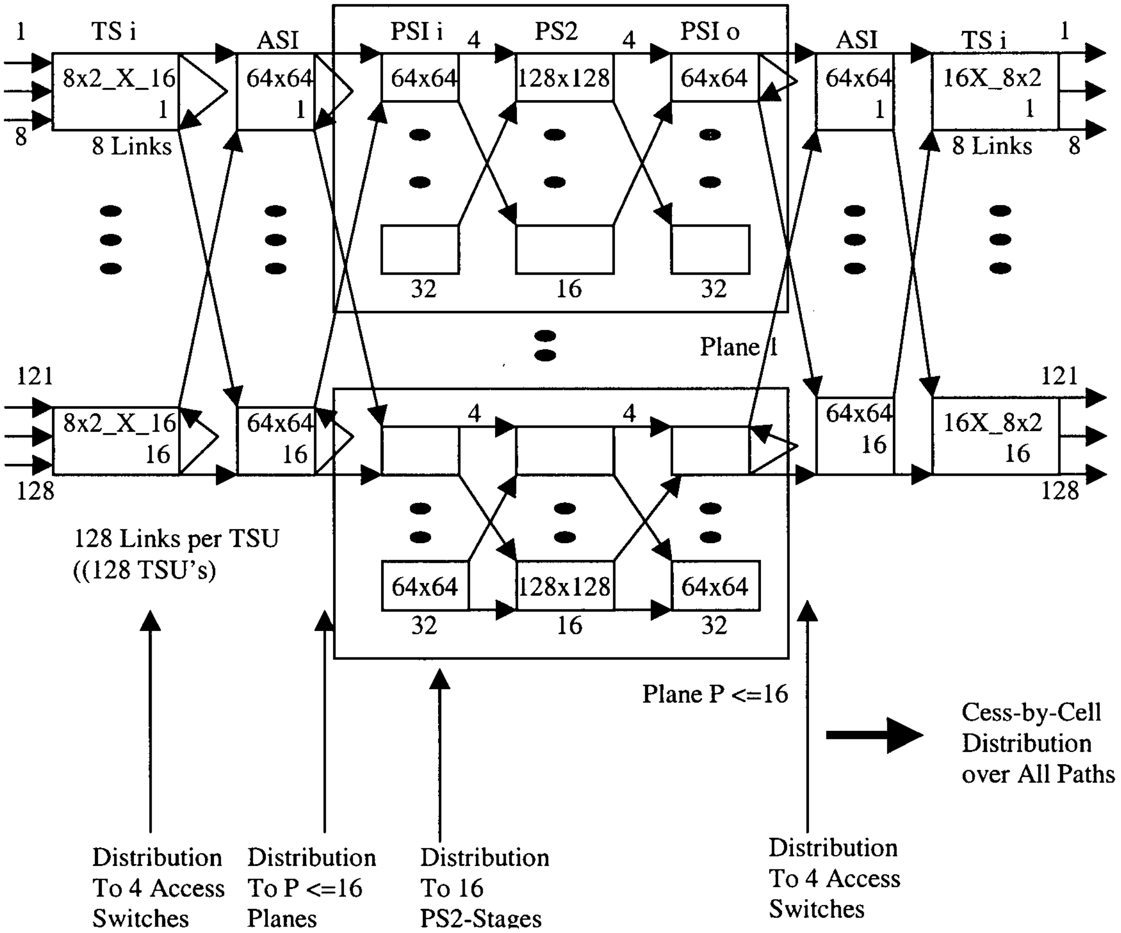


Figure 20 : Roxanne Switching Fabric [22]

The switch shown in the above figure has 128 links on each terminal subscriber unit (TSU). Physically, 4 input links with the capacity of 150 Mbit/s are demultiplexed from an input cable of 600 Mbit/s.

The basic block of the switching fabric is the Roxanne Switching Element described in Section 3.2 with 128 inlets and 128 outlets. This basic block is used in ASI, ASO, PS1I, PS1o and PS2. ASI and ASO are actually in the same block where ASI uses half of the block and ASO uses the other half. The same architecture is used in Ps1i and PS1O.

On each TSU, there are 16 terminal module (TS) and 4 access switches (AS). The TS distributes cells over the 4 AS randomly. Since each TS has 16 outputs but only 8 inputs, the internal load is half of the external load. Then, cells are distributed on the planes and distributed again from PS1I to PS2. The distribution process is used to ensure the maximum utilization of the resources and minimize the occurrence of congestion. Direct routing will be performed in PS2 and AS0. Under this configuration, even if some of the switching element fails, they will not affect the whole system seriously. For example, if one of the 4 AS fails, the internal load will only increase by 4/3.

Failure Detection

In order to minimize the effect of the faulty element, error detection must be done continuously and identify the fail switch in the shortest time. The Roxanne Switch has a self-checking routine which continuously sends a "Backward Availability Signal" from the last stage of the module to the first stage. A switch module will receive the signal if its next stage functions properly. If it cannot receive the signal, the next module (or the module that follows it) must fail so it will not transmit any cells to that faulty module.

Resolving Out of Sequence Problem

The Roxanne Switch has a special method to deal with the out-of-sequence problem. When a cell enters the network at one of the TSi, it is time stamped. When it leaves the network at TSo, its time is found again. If the visiting time of the cell in the switching network is less than the maximum delay in the network, then it will wait in the output until the time that the cell has spent in the network equals the maximum delay. Therefore, each cell spends the same amount of time in the network and the cell sequence is guaranteed. However, there will be a constant (longer) latency in the switch, but the throughput is maintained.

2.2.2.3 MIN without Internal Blocking

2.2.2.3.1 Batcher-Banyan

The Batcher-Banyan is one example from the class of MIN without internal blocking. The assumption is that cells may not be destined to the same outlet at the same time. Therefore, special arbitration logic and a buffer have to be implemented at the entrance of the switch fabric. Therefore, head-of-line blocking may be present at the entrance of the Batcher network.

The Batcher-Banyan is composed of a Batcher Network and a Banyan Network as shown in Figure 21 [22].

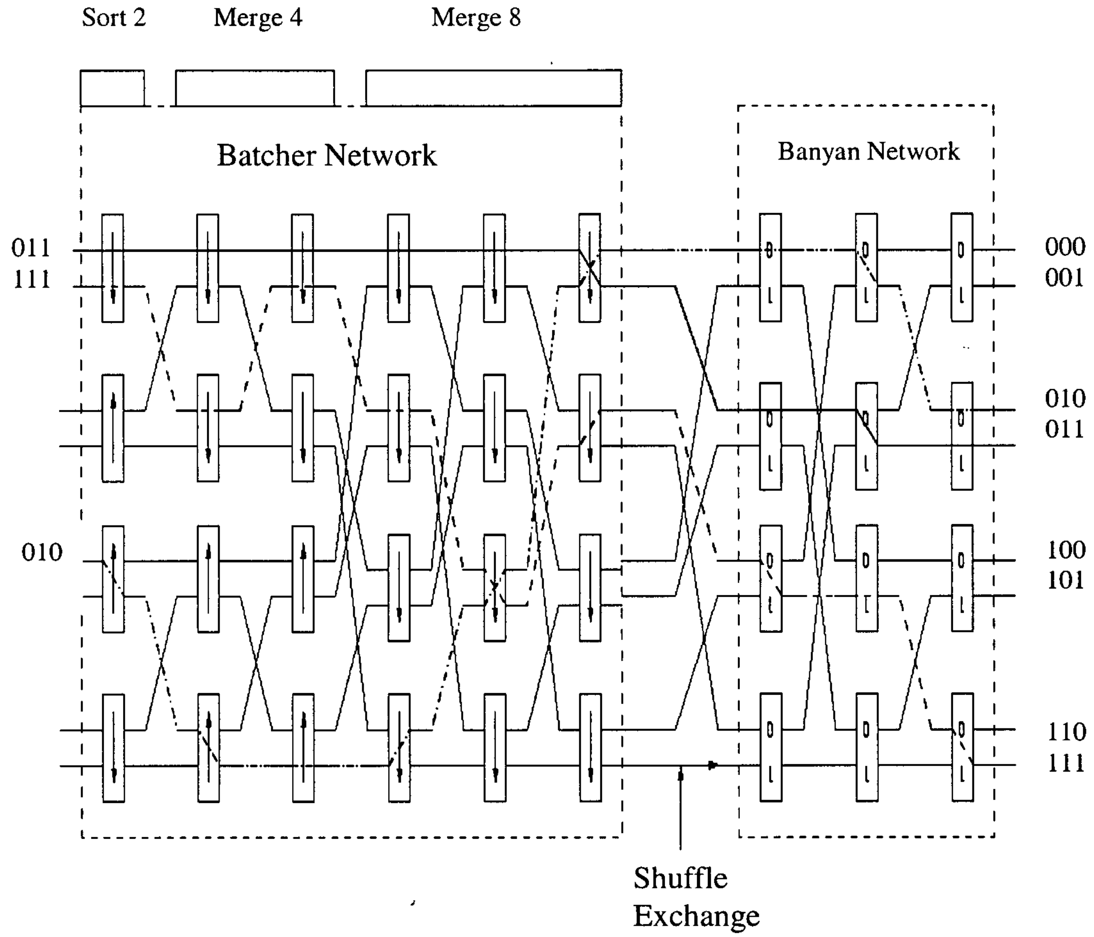


Figure 21 : Batcher-Banyan Network Topology

The way the two networks are interconnected is called Shuffle-Exchange.

The Batcher network is a sorting network which sorts the incoming cells according to their destination address. In this example, the smaller address will be routed to the higher outlets, while the bigger address will be routed to the lower outlets. All the output cells will be put on the topmost outlets. The arrow on each sorting element points to the outlets at which the largest address will be routed. This sorting helps to avoid internal blocking inside the Banyan network.

The output of the Batcher network is inputted to the Banyan where the cells are routed to their specific outputs. The Banyan network is a self-routing network, that is, the routing information is stored in the tags of the cells. It is guaranteed that no internal blocking will occur if the incoming cells are properly sorted.

2.2.2.3.2 Starlite

Proposed in 1984, the architecture of the Starlite switch is based on the Batcher-Banyan network. The switch can handle fixed length cells. A routing tag is added to the cells entering the network. Due to the Batcher-Banyan network, Starlite is very suitable for VLSI implementation because of its regular structure. It uses the recirculation buffering approach to handle output contention. The architecture of the Starlite Switch is shown in Figure 22.

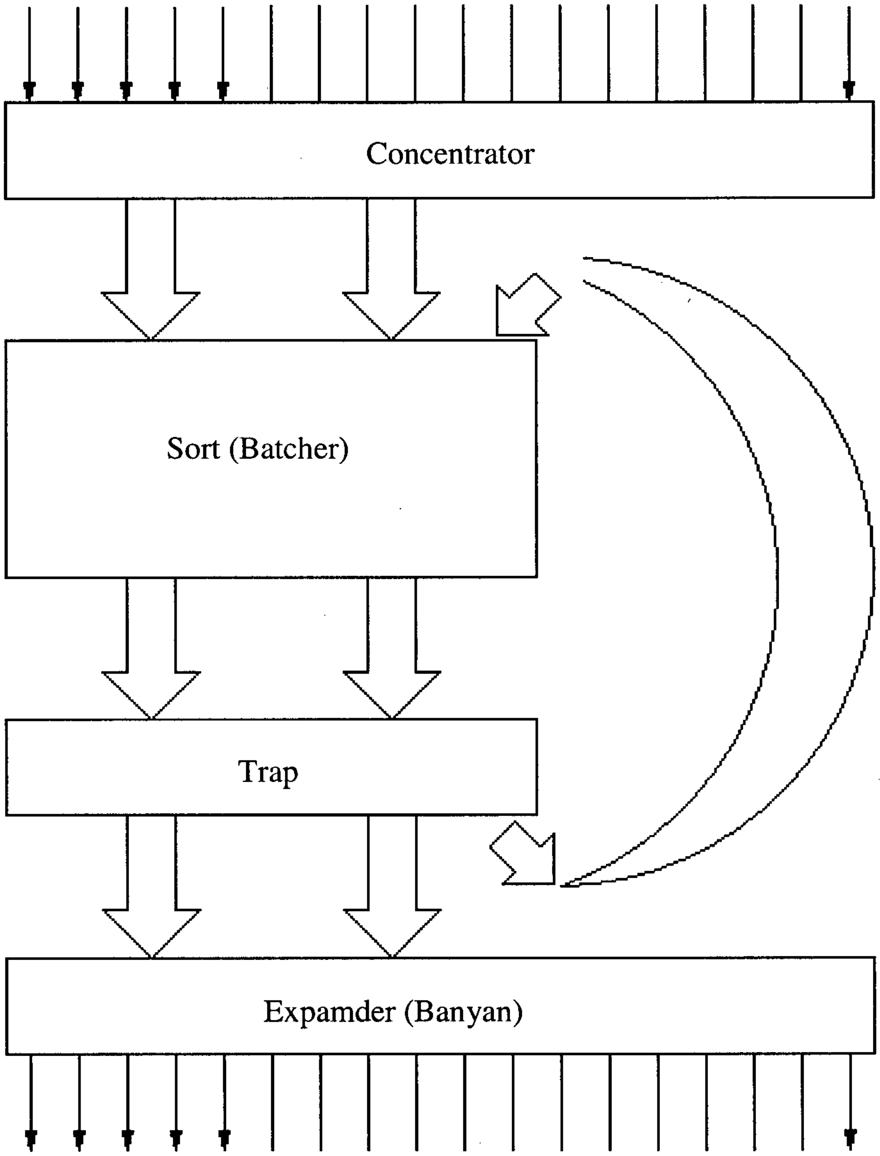


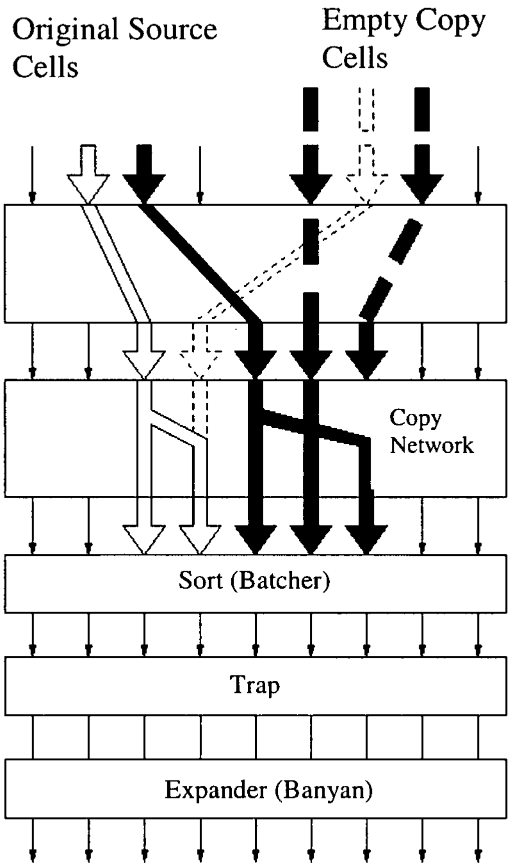
Figure 22 : Starlite Switch Architecture

The first part of the network is a concentrator. As with the concentrator in a Knockout Switch, some cells may be dropped in the concentrator, but the probability should be very low. There is an activity bit associated with every cell, and that bit indicates if a cell is empty or not. By using the running sum of the activity bit, the concentrator will direct the non-empty cell to the connected outputs. Cells arriving at the non-connected output will be dropped.

Cells which successfully pass through will enter the Batcher network and will be sorted at the output. The most interesting occurrence transpires in the trap network. It detects duplicated destination addresses (those that will cause an output contention problem) and moves the cell with a duplicated address to the right of the output of the trap network, except for the first instance of the cells with the duplicated address. Since the Batcher network will group cells with the same address together at its output, only one stage of comparators is required to detect any duplicated address as in the trap network. Then the cells with the duplicated address will be fed back to the Batcher network. There is a buffer at the input of the Batcher network for the recycled cell's because there may not be enough free input at the Batcher network. Since there may be another newly arriving cell with the same destination address as the recycled cell in the Batcher network, the recycled cell may be recycled again. In order to avoid any out of sequencing created in this situation, the recycled cells are indicated as "aged" so that they have higher priority in the Batcher network. There should be a bit in the header that can indicate the cell is aged.

2.2.2.3.3 Multicast/Broadcast Capability of Starlite

In order to perform a multicast/broadcast function, a sort-to-copy network and a copy network is added on top of the Batcher network. This is shown in Figure 23.



ACTO	Source Address	COPYO	Destination Address	Info

Original Source Cells

ACTO	From-Who to Copy	1	Self Address	Empty

Empty Copy Cells

Figure 23 : Multicast in Starlite

The copy network uses the empty copy cells to make copies. The copy cells have a structure similar to the normal cell, except that they do not contain any data and they have "1" in the copy bit. The sort-to-copy network will group copy cells, and their original, physically on adjacent lines. Then, the copy network will copy data into the next the adjacent cell until the cell has a different source address.

Chapter 3 Bandwidth Optimization in Banyan Switches

3.1 Introduction

The internal blocking of the Banyan network can be improved by placing internal buffers and increasing the internal link capacity as was seen in Chapter 2. The question of how much internal buffer and the internal link capacity are required depends on a number of factors including:

1. The extent of performance improvement in term of cell loss, delay, and throughput/goodput: Feasibility and cost of implementation which includes such factors like interconnection and I/O density, switch modularity, and growability.
2. Impact on QoS and cell sequence.

In this chapter, we first proceed to examine these crucial factors in some detail.

3.2 The Buffered-Dilated Banyan (BDB) Switch

The basic architecture of the BDB Switch is shown in Figure 24. It consists of a multistage switching fabric, with each stage consisting of switch elements connected to dilated links. A switch element contains three components: input ports, output ports, and output buffers (or buffered concentrators). Each port has different multiple input and output links. The links on both the input and output port may not be equal; and therefore we called it the Dilated Switch Element (DSE). Figure 24 illustrates a KxK DSE with a dilation of M at the input ports and L at the output ports. Earlier work has shown that such a DSE can have a regular structure when constructed from smaller basic switch elements [11,16]

A dilation of M on an input port indicates that M links are feeding a single input port. Similarly, a dilation of L at an output port indicates that such a port is capable of transmitting up to L cells over L links simultaneously. Since up to KM cells can be received and forwarded simultaneously to an output port, a buffer is required to queue the cells and to concentrate them onto the L output port. Variable dilation per stage is used for optimizing the internal bandwidth as well as to meet a specified cell loss probability. In the first few stages of the switch, the number of input and output links will grow from one stage to the next to avoid link contention. However, the number of links will either remain fixed or decrease in the last few stages.

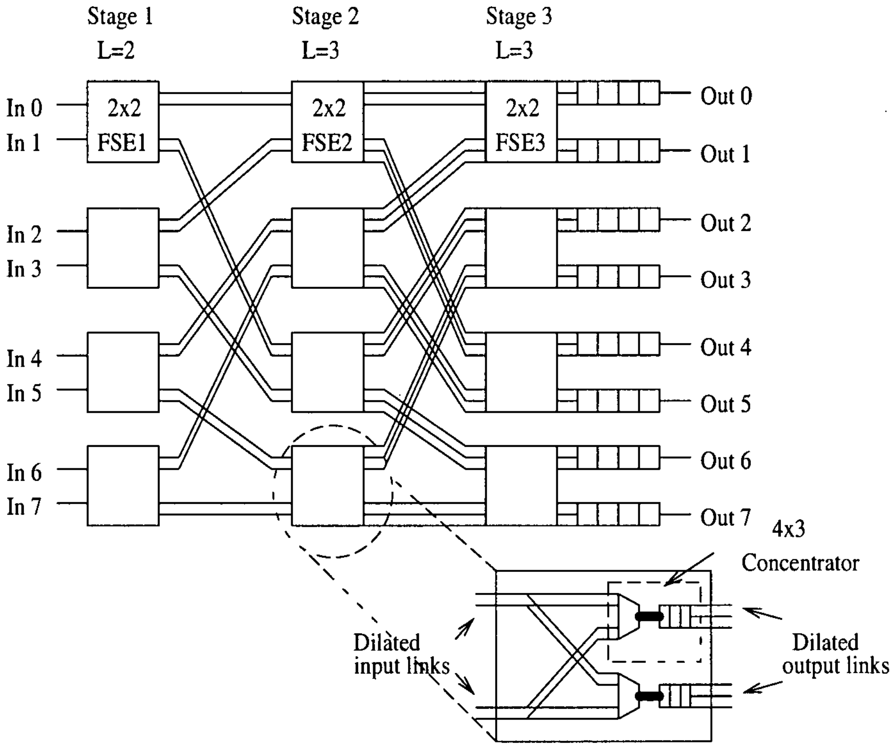


Figure 24 : 8x8 Banyan Switch Architecture

3.3 BDB Architectural Features

Switches perform three major functions: cell routing, cell buffering and back-pressure controlling. Routing, or cell distribution, is done over an interconnection medium that normally includes memoryless cross-interconnect switching elements, (SEs) or shared-medium modules. It is important to note that buffering in a switch serves two distinct purposes, depending on whether it is employed in the switch fabric or at the switch ports (see Figure 24), as explained below:

1. In the switch core, small amounts of strategically placed simple (FIFO) buffers can be used control cell loss and to enhance switch robustness for changing traffic characteristics.

2. The major switch buffering resources are located in the line-card modules at the input and output ports. These are larger and more complex per-VC buffering structures, which are used to control rate allocation, guarantee fairness, and perform real-time scheduling of cells on the output links.

The proposed switch only employs distributed back-pressure to eliminate cell loss from the internal stages of the switch, and to limit cells to the BDB switch inputs. The back-pressure mechanism has several advanced features that are described in Section 2.1.3. One particular feature is the use of a window forwarding policy in conjunction with back-pressure to eliminate head-of-line blocking in the DSE buffers. It is worth noting that back-pressure signals can be generated either by congested output-port buffers or by congested DSE buffers.

The sequential operations of a switch are enqueuing, cell forwarding, and back-pressure controlling. A detailed description on the operation of each component is presented in Sections 3.3.1 to 3.3.4.

3.3.1 Enqueuing in the DSE

When an incoming cell arrives either from the input buffer or from the previous stage, the enqueuing operation immediately takes over responsibility. Enqueue functions to check the destination port addresses in the arriving cells. The corresponding number of bits (logK/log2) from the destination port address will be selected to determine the appropriate output buffer, that is, in the first stage of DSE, the number of bits will be selected from the first logK/log2 bit. For K=2, the first bit in the destination port address will be used to determine which buffer is suitable for the cell to enter or store. If it has a value of 0, the cells will queue into the upper buffer (see Figure 24)., and if the value is 1, it enters the second or the bottom buffer. In the second stage of DSE, similar to Stage 1, the second bit from the destination port address is used to determine which buffer is designated for the cell to enter next, and so on.

3.3.2 Cell Forwarding in the DSE

After the enqueuing operation, the next step of SE is cell forwarding. The basic principle of cell forwarding is to send out all the cells from the buffer until it reaches the end of queuing, or the output dilation L links are all occupied on the same simulation cycle. An HOL blocking may occur when there is a received back-pressure signal from the previous cycle. A back-pressure signal is used to indicate a buffer congestion on a switch element of the next stage. HOL blocking causes a negative effect on the performance of the switch core. Many different schemes, which can solve this problem, were already described [13]. The most simple and common scheme to solve HOL blocking is a window scheme (or called a sliding window). A window scheme skips the blocking cell in the queues, then searches for the next cell and continue to perform the cell forwarding. A detailed description on the sliding window is described below.

A window scheme (sliding window) must co-operate with back-pressure control. The sliding window feature is applied whenever there is a back-pressure (BP) signal received from the next stage of the previous cycle. During the cell forwarding operation, the FIFO buffer sends out the first buffer to the queue if there is no existing BP signal. If a BP signal is present, the sliding window feature will be used to compare the BP signal info with the destination port address inside the cell. Verification will be done based on the destination port address of the cell and the status of the back-pressure signal from the next stage. If the verification is valid, the sliding feature will keep the HOL cell in the queue, and then look at the next cell until it reaches the end of the queue, or all L output links are fulfilled. Without a window scheme, if the cell matches the BP signal info, an output link will be idle for the present cycle. One of our current research projects completed the physical layout of multi-FIFO, that is, being capable of operating at 5 Gbit/s. This extremely fast response can be used for implementation of sliding window.

3.3.3 Backpressure (BP) Controlling

Backpressure controlling is a technique which is used to eliminate cell loss in the output buffer, and to control the cell loss within the switch fabric. This is called link-capacity (LC) backpressure. Nevertheless, the cell delay increases. When the number of cells stored in the buffer on the SE reaches a preset buffer threshold (BT), a backpressure signal is generated and sent to all input ports. The input ports will then send the backpressure signal back to the corresponding output port on the previous stage (see Figure 9).

For a special situation when congestion occurs in a buffer on the first stage, the backpressure signal is sent to the appropriate input buffer (see Figure 25). Similarly, if an output buffer is congested, a backpressure signal is sent to an appropriate output port in the last stage (see Figure 25).

Figure 25 showed an example of how backpressure is implemented on an input-output buffered switch with a memoryless internal switch buffer. In the example, six HOL cells from input buffers 0,2,4,5,6 and 7, are destined to output buffer 0, but the output dilation of the last stage is 3. The HOL cell from input buffer 7 is blocked by the second SE at stage 2, and the HOL cells from inputs 5 and 6 are blocked by the top SE at stage 3.

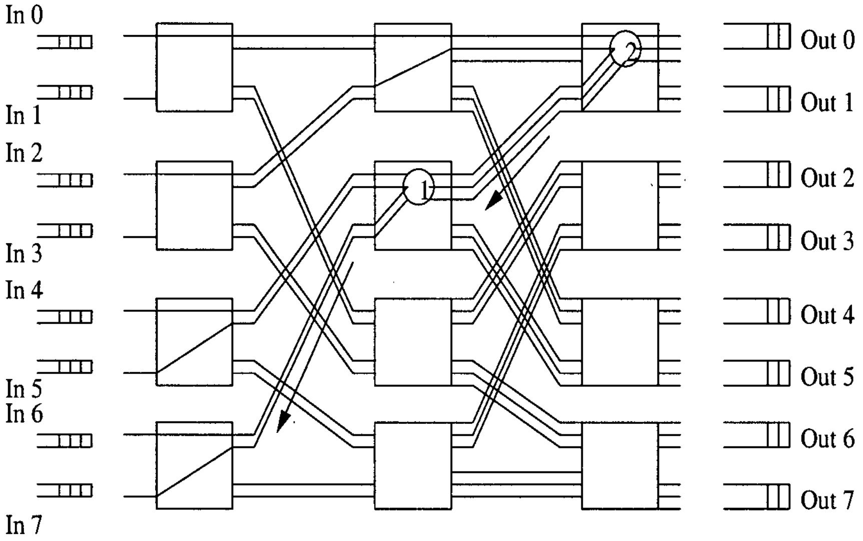
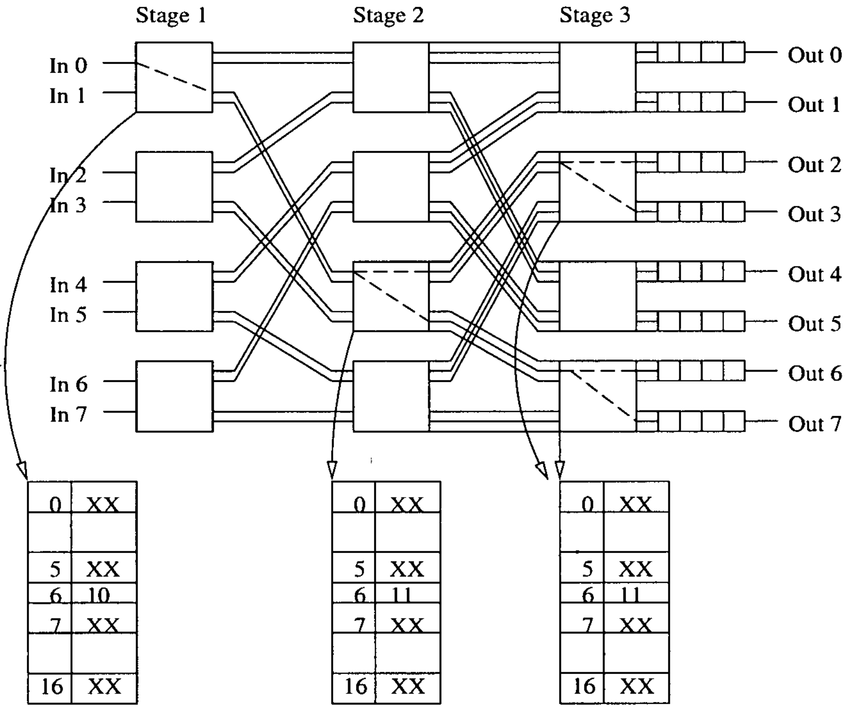


Figure 25 : Concept of LC Backpressure in Input-Ouput Queued BDDB with Internal BS

With an additional internal switch buffer size of BS j, j is the number of stages of the SE. An identical six HOL cells from thė above example are showed in Figure 25. The HOL cell from input queue 7 is stored onto the top internal buffer on the second SE at stage 2. If the internal buffer size exceeds the preset threshold BTj, a backpressure signal will be generated and propagated on a reverse path back to the previous stage (i.e. the upper buffer on the bottom SE at stage 1), and so on. Similarly, the HOL cells from input queue 5 and 6 are stored onto the top of the internal buffer on top SE at stage 3. Under the same circulation, if the buffer size exceeds the threshold value BT, a backpressure will be generated and passed back to the previous stage.

3.3.4 Multicasting Capabilities

Our switch fabric has excellent multicasting capability and the design is based on IgT technology [14]. Figure 26 illustrates the routing of a multicast connection from port 0 to port 2,3,6 and 7 in a 8X8 FAB. Each DSE consist of an internal multicast group table. Each of the entries of this table correspond to a multicast group and are used to determine the number of cell copies needed, as well as the destination of each copy cell. The cell duplication is performed inside the enqueing function. Multicast cell information update occurs during the cell forwarding function. Figure 26 shows an example of point-to-multipoint (multicasting). A multicast cell with a different output destination must be checked and assigned a specific group number. In our example, a group number 6 is specified to this multicast cell from input queue 0 to output port 2,3,6 and 7. If a match does not exist, a new group number can be added to the multicast table on each SE easily. The traditional way to update all multicast tables is to prepare a special cell with a new group number, and all specified destination port addresses, and place them into all input ports once the special cell enters a SE. The SE will update its own multicast table based on the destination port address listed on the special cell, which is similar to distributing an incoming cell to an appropriate buffer during cell forwarding.



0	XX
5	XX
6	10
7	XX
16	XX

0	XX
5	XX
6	11
7	XX
16	XX

0	XX
5	XX
6	11
7	XX
16	XX

Figure 26 : 8x8 Point-to-Multipoint Connection

Chapter 4 Buffer Management

4.1 Queuing Technique

4.1.1 Introduction

In order to solve the contention problem, there are four types of buffering strategies. Input queuing, output queuing and central queuing are applicable to switching elements. The recirculation buffer is used as an external buffer in a switching fabric.

4.1.2 Input Queuing

In this method, each input link of a switching element is provided with an input queue, as shown in Figure 27. The queue may be a FIFO queue (round robin fashion). The arbitration logic is used to control the flow of cells into the switching transfer medium by determining if there will be any internal contention problem or output contention problems arising after a cell is dispatched from the queue. If there is no contention hazard, the arbitration logic will remove a cell and put it on the switching transfer medium which will transfer that cell to a specific outlet.

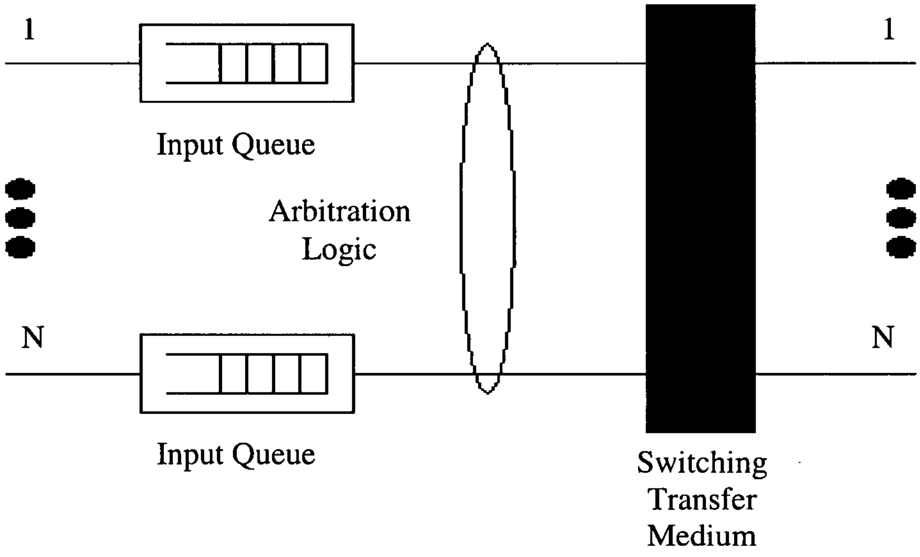


Figure 27 : Switching Element with Input Queues

However, there is a blocking problem with this technique. When the arbitration logic scans a cell in the head of the queue and finds that the cell must wait in the queue,. the other cells behind it must also wait even if the cells can be served without contention. This is known as Head of Line (HOL) blocking.

4.1.3 Output Queuing

Unlike the input queuing solution, the output queuing approach allocates one output queue on each outlet of a switching element. The architecture is shown belowin Figure 28 [22].

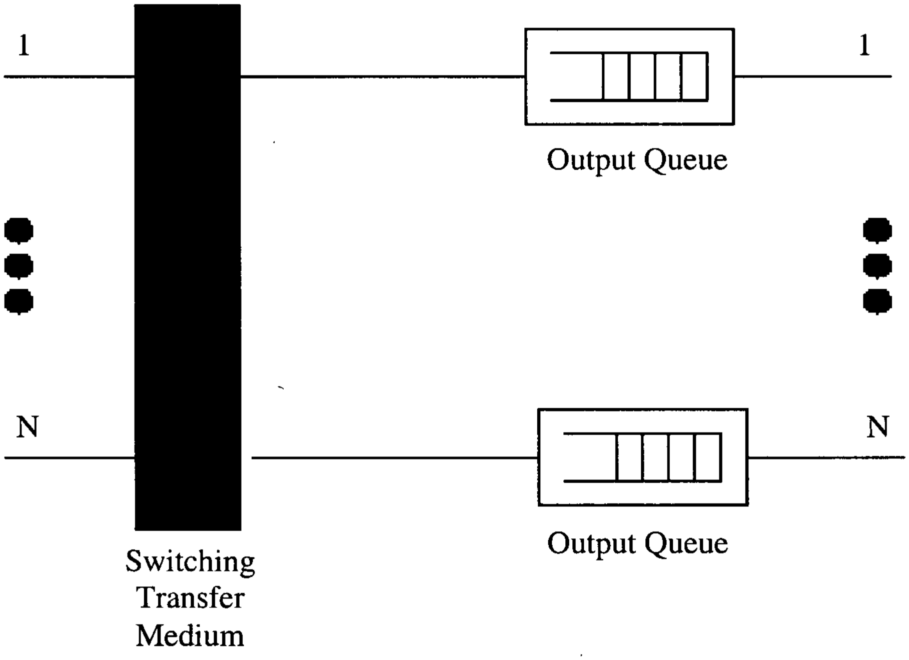


Figure 28 : Switching Element with Output Queues

This solution solves the output contention problem without the HOL blocking problem in input queuing. In addition, no arbitration logic is need. The only requirement of the output queue is that that it must ensure the correct sequence of the cells waiting at the queue. A simple FIFO queue can satisfy that.

Since cells may arrive at the input of the switching transfer medium at the same time, the medium should transfer cells at N times faster than the speed of input in order to ensure no cell will be lost. Therefore, N cells should be transferred to the N output queue at one cell time. Moreover, N cells may want to access one particular outlet. Thus, the output queue must be able to perform N writes (plus 1 read to send out a cell) in one cell time.

4.1.4 Central Queuing

In the central queuing approach, queuing buffers are shared between inlets and outlets. Proper memory management logic maintains the free locations in the buffers for the arriving cells, and should determine which cell should be put on the output. Therefore, the central queue may be accessed randomly; however, the sequence of the cells must be maintained. At its maximum performance, the central queue can perform N reads and N writes in one cell time. Thus, the complexity of the queuing buffer and the memory management logic will be greater than the previous two approaches. Figure 29 shows a block diagram of central queuing.

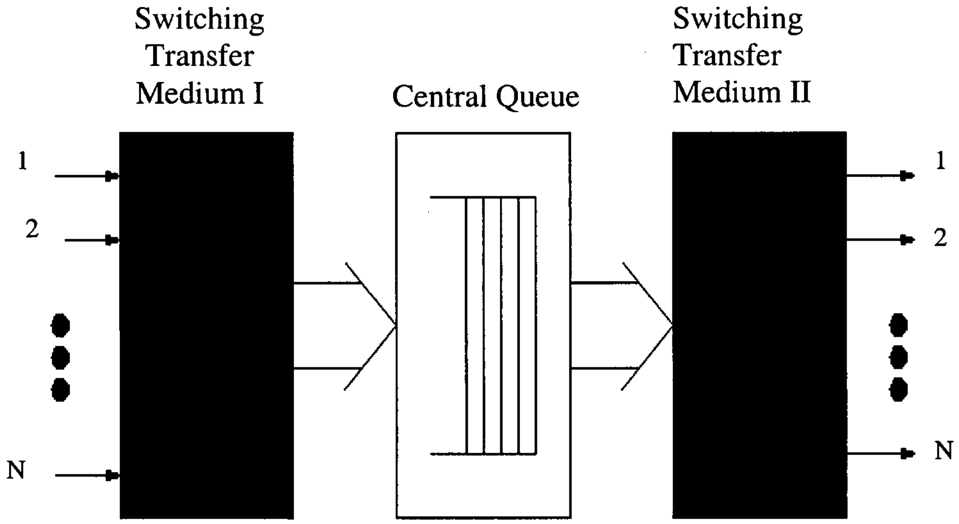


Figure 29 : Switching Element with Central Queuing

4.1.5 Recirculation Buffering

This is an external buffering approach. If cells cannot be sent to the output port due to output contention, they will be fed back to the input via a set of recirculation buffers. However, this approach may cause cells arriving at the output to be out-of-sequence and special logic must be used to avoid that. This performance of this approach is similar to the output queuing technique.

4.1.6 Comparison of Different Queuing Approaches

Queuing Time

In terms of queuing time, under the same load, input queuing will have the longest queuing time due to HOL blocking. Central queuing and output queuing will have similar queuing times.

Load

In terms of the impact of load on each queuing approach, the maximum load of input queuing is 58.6% if N (number of inlets) goes to infinity. Figure 30 shows how input utilization (load) affects the average queuing time for an output queuing system For output queuing, the average waiting time is quite small under a load of 80%. For loads more than that, the average waiting time increases exponentially.

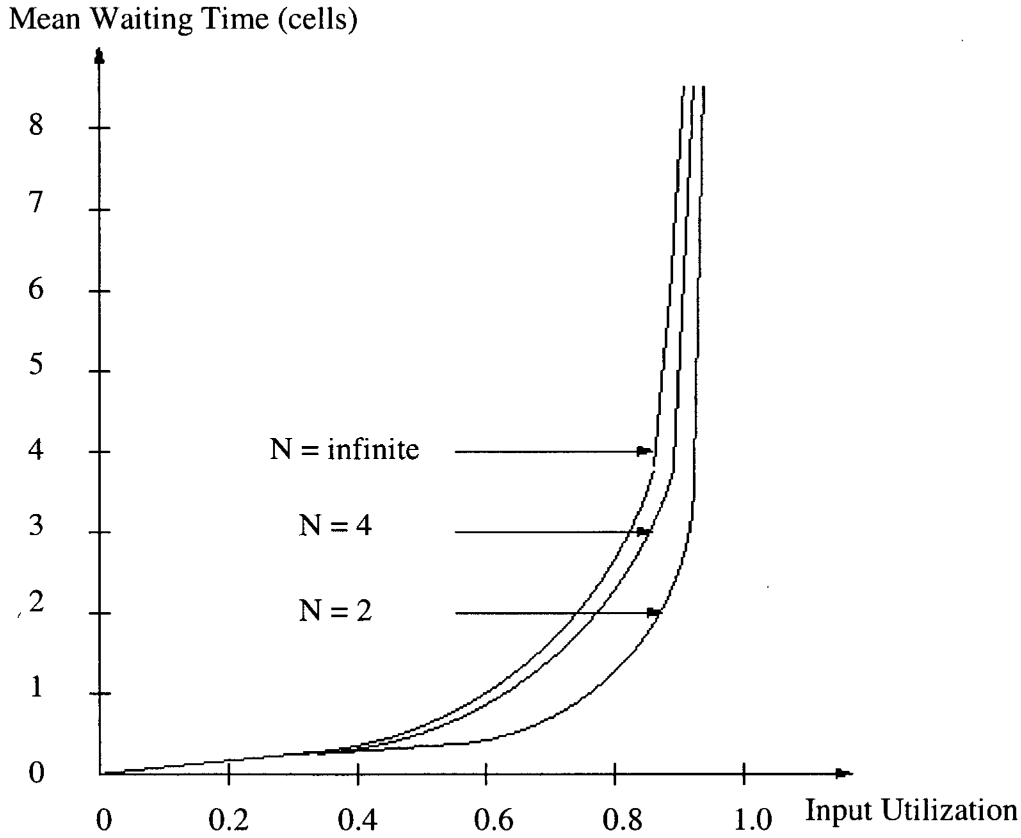


Figure 30 : Mean Waiting Time for Output Queues

Buffer Space

The central queuing approach has the smallest buffer space since not only can the input queues and the output queues be shared, but the queues on different outlets/inlets can be shared too. Therefore, memory space can be used more efficiently. As a result, the required buffer size of central queuing may be smaller than the other queuing approach for a given cell load and probability of cell loss. Figure 31 compares the queue size of different queuing approaches under different load and at the cell loss rate = 10^{-3}. The 10^{-3} loss rate is actually too high to be realistic. This loss rate is chosen because it shortens the time for computer simulation to obtain the result.

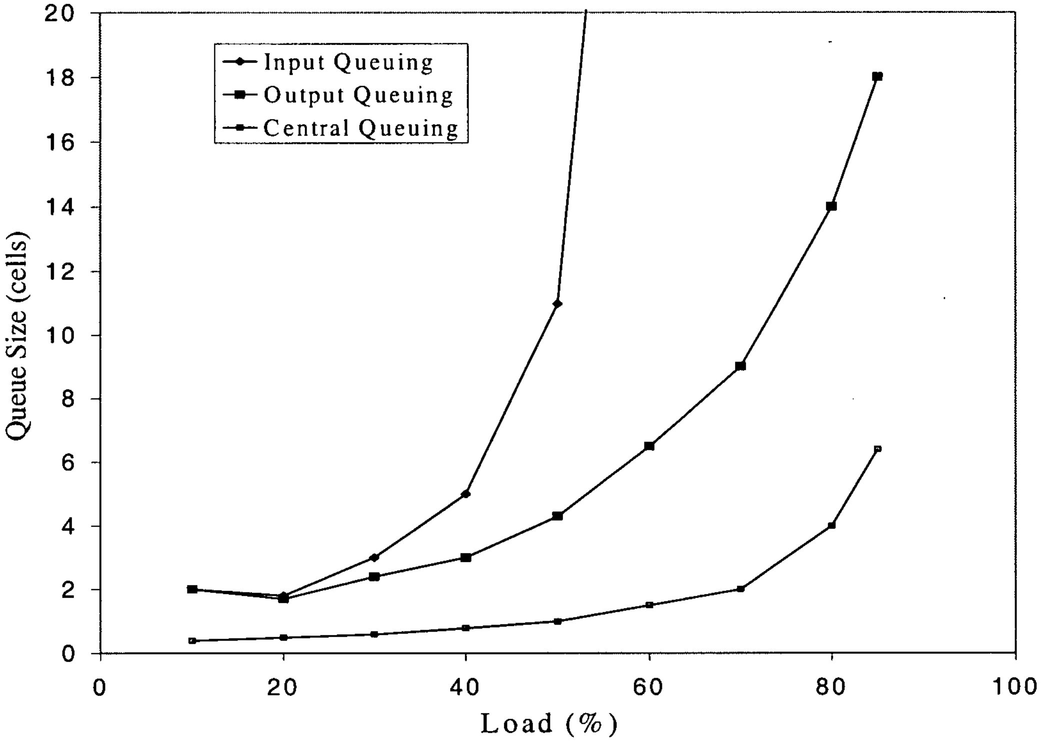


Figure 31 : Queue Size as a Function of the Load (Cell lost rate = 10^{-3})

Access Time Requirement

In terms of the access time requirement in the buffer, central queuing has the strictest access time requirement. this is because every inlet and outlet may access the memory at the same time so that reads and writes may be executed together in one cell time. The access time for input queuing may be longer since each queue only deals with one inlet so that only one read operation will be performed in one cell time. The access time of the output queue is between the one of input queuing and central queuing. Table 4 shows the required memory access time for the three queuing approaches for both single ported and double ported memory.

	Input queuing	**Output queuing**	**Central queuing**
Single ported memory	W/(2F)	W/((N+1)F)	W/(2NF)
Example	106.7ns	6.7ns	3.3ns
Dual ported memory	W/F	W/(NF)	W/(NF)
Example	213.3ns	6.7ns	6.7ns

Table 4 – Requrired Memory Access Time

Data used in the example:

Width of data bus(W) = 32bit

Link speed (F) = 150Mbit/s

Number of inputs(N) = 32

4.2 Buferring Management

4.2.1 Introduction

Traffic control is a key word in network stability. One can broadly distinguish two kinds of control: preventive and adaptive. Preventive control is based on the notion of a traffic contract where sources adjust their sending rate to fit in a control scheme. The network in turn, gives QoS quarantees. Adaptive control is intended to use the available bandwidth. It is generally suitable for applications not requiring QoS commitments, and can be classified as a "Best-Effort" service class. Adaptive control can be based on explicit feedback signals from the network (e.g. ABR service class) or on network response to variations of source behavior. As one can easily note, TCP/IP is the immediate transport layer candidate for today's applications. TCP uses an adaptive window-based flow control. The congestion avoidance and control algorithms deployed by TCP aims at using the available network bandwidth.

Random Early Detection (RED) was introduced by Floyd and Jacobson. This technique tries to keep the average queue size as low as possible, while allowing occasional bursts [24]. The authors show how RED can maintain high throughput while minimizing delay. The idea is to monitor the average queue size and hence avoid packet drops when the network load changes. RED can also identify connections that use a large share of the total bandwidth.

The simplest form of packet discard, called Drop Tail, discards arriving packets when the input port buffer space is exhausted.

4.2.2 Random Early Detection (RED) Technique

Feedback from the network provides the means for applications to identify the network state and to monitor their sending rates. Feedback can be used in packet networks in two ways: implicit and explicit. Explicit feedback can use a special field in a packet (cells for ABR) to indicate congestion. Implicit feedback is based on network response to the variation of the source behavior. It is usually deducted from delay variations of packet acknowledgements or packet losses.

RED Algorithm

RED uses an implicit congestion notification by means of packet drop. Rather than waiting for the queue to become full and start dropping each new arriving packet, RED decides to drop arriving packets with a drop probability triggered each time the average queue size exceeds a certain threshold.

RED maintain an estimated average queue size using a negative exponential weighted moving average [23]. This estimated queue size is used as the basis for probabilistically dropping or marking packets. A parameter, $weight_q$, is used to control the rate at which the estimated average queue size reacts to changing network conditions. This parameter must be carefully set for optimal network performance. If it is set too low, then the RED will react too slowly and cause sustained congestion. On the other hand, if $weight_q$ is set too high, then the RED will react too quickly and respond inappropriately to transient congestion, resulting in underutilization.

The low-pass filter used to calculate the average queue size ensures that short-term increases in queue size due to the transient congestion or bursty traffic, do not significantly increase the estimated average queue size. The filter used to calcuate the average queue size is an exponential weighted moving average given by the following equation:

$\text{AvgLen} = (1 - w_q)\ \text{AvgLen} + w_q * \text{queue\_size}$ [23];

where w_q is a constant satisfying $0 <= w_q <= 1$ and the queue_size is the queue size.

There are two thresholds to which the RED compares the average queue size: Threshold_{min} and Threshold_{max}. If the calculated average queue size fall below the Threshold_{min} parameter, no packets are dropped. However, when the average queue size falls between the parameters Threshold_{min} and Threshold_{max}, the arriving packets are dropped with a probability that is a function of the average queue size and packet size:

P_a=P_b/(1-count*P_b) and

P_b=max_p* (AvLen-$Threshold_{min}$)/($Threshold_{max}$ - $Threshold_{min}$)

P_b=P_b * Packetsize / Maximum Packet size

where count is the number of queued packets as long as AvgLen remains between the two thresholds. It is set to zero for each loss. Max_p is the maximum dropping probability, and is set to 0.02 on all our simulation [23][24]. The above equation also states that a large packet is more likely to be dropped than a small one.

Chapter 5 Performance of the Packet Switches

5.1 Introduction

The proposed switch was explained in Chapter 4. In this chapter the simulation traffic model is introduced first, followed by the simulation data collected for the proposed switch.

5.2 Input Traffic Model

From Sections 5.3 to 5.9, there are simulation results and analyses on two different traffic loads and conditions, as presented the following.

5.2.1 Uniform Traffic Pattern

The first traffic pattern to be described and tested, is the independent uniform traffic pattern. The independent uniform traffic pattern refers to a traffic pattern in which the cell arrival at the input is based on the Bernoulli process, and the destinations of the cells are independent of each other and uniformly distributed.

5.2.2 Bursty Traffic Pattern

Bursty traffic is the second traffic pattern used in the simulation. A bursty pattern refers to traffic that has active and idle period during a transmission. Its model is considered an "open loop" [15]. The bursty period is generated at a maximum rate regardless of the congestion in the network. It consists of three parameters: Bursty Length (BL), Idle Length (IL), and Destination Address. During an active period (bursty length), the parameter BL is determined independently and fall into a negative exponential distribution with a mean value called the Average Burst Length (ABL). When ABL is equal to 15, the maximum value of a bursty length can reach up to 291. It can reach up to 525 when the ABL is at 30. Idle length is a parameter for the idle period. The mean value of idle length can be determined in terms of the active period and the load using the equation $Mean\ of\ Idle\ Length = Active\ Period * (1-p)/p$, where p is the load. The destination address is randomly generated by a uniformly pseudonumber generator. For a bursty pattern, an input port receives a fixed size ATM cell with the same final address during the active period. After the active period is terminated, the input port is turned into idle until the idle period is completed. A new bursty pattern occurs after both the active and idle periods.

5.2.3 Load Varies under Poisson Distribution

In order to meet the minimum requirement of all the existing switch cores, the simulation must be at a high fixed load rate. These results are shown in next section. To be more realistic, the load is varied every 1 million simulation cycles over 30 million cycles, and all the loads are under Poisson distribution with a parameter called the mean average load (Avg_load). Figure 32 shows the performance of the load under Poisson distribution over 30 million cycles, with a mean of 0.8. All the simulations under Poisson distribution are illustrated in Figure 32.

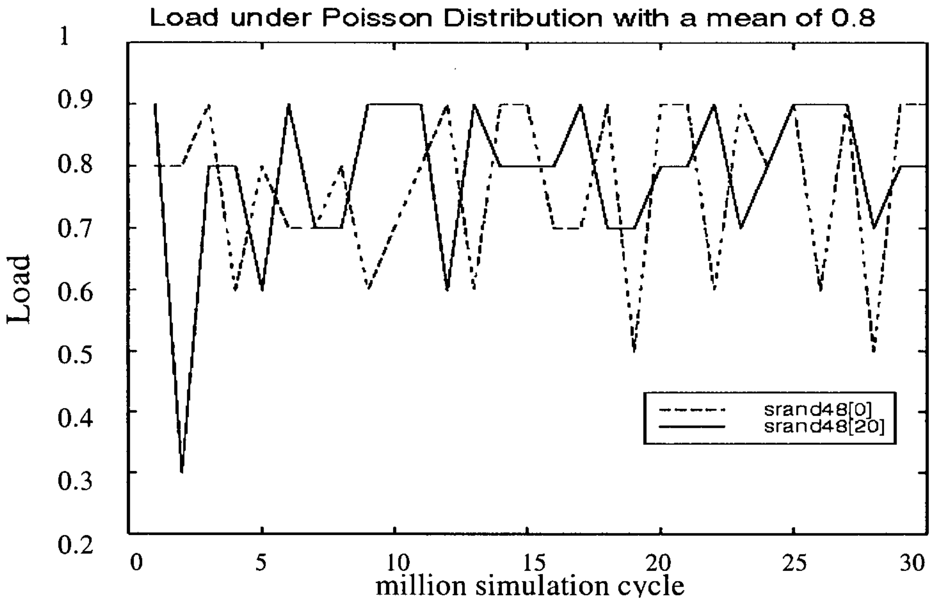


Figure 32 : Load under Poisson Distribution

5.3 Performance Analysis under Independent Uniform Traffic

In this section, the advantages of the proposed switch, which is based on the congestion control mechanisms as discussed in Chapter 3, are demonstrated using computer simulation. Simulations were carried out using 95% confidence intervals for uniform (Bernoulli), as well as bursty traffic sources (see Section 5.4). A simulation was performed on the proposed switch with different dilation and buffer-size parameters. The results, which are shown in Figures 33 to 57, demonstrate superior performance in packet loss probability under a uniform traffic pattern. Figure 33 to Figure 49 represent the circumstances where the switch is handling ATM traffic. Additionally, Figure 50 to Figure 57 illustrate the conditions when the switch is handling IP traffic.

Figure 33 shows a dramatic decrease in cell loss probability through small increments of the internal buffer size (BS) in the DSEs. This strongly indicates that internal buffering significantly enhances the performance of cell loss probability in the switch core.

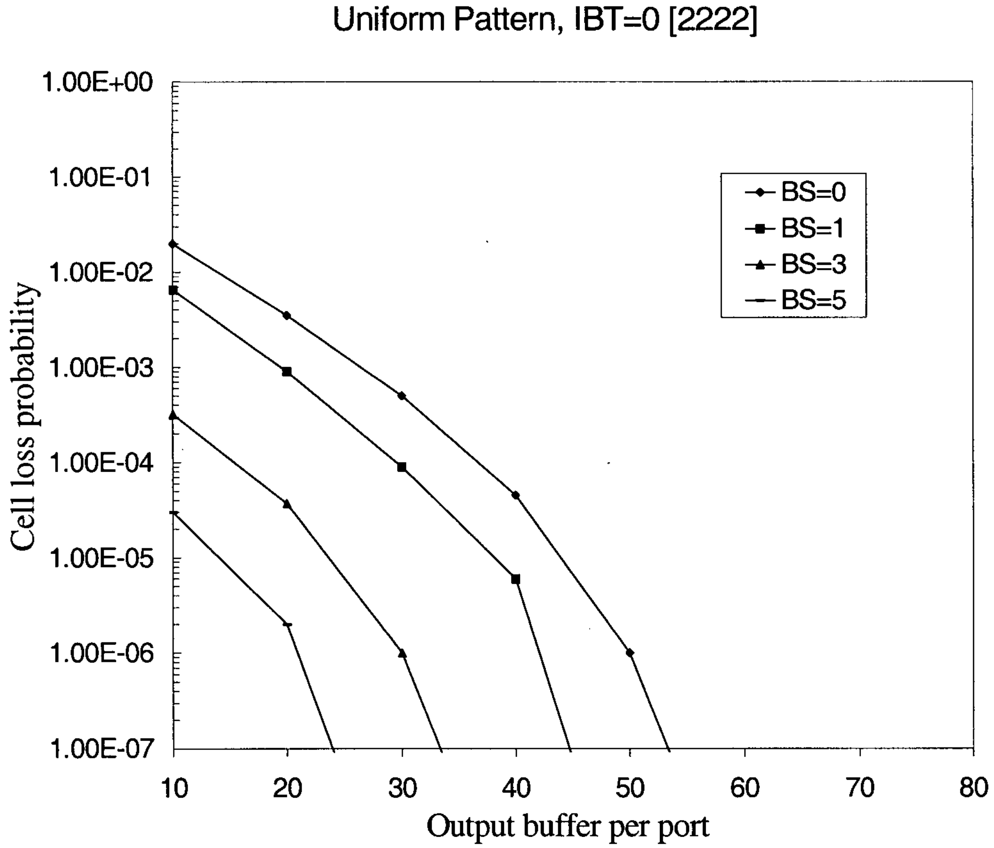


Figure 33 : 16x16 Banyan Switch under Uniform Traffic with 0.95 Load

5.4 Cell Loss and Stage Dilation

In order to verify our non-blocking switch core, the first simulation result is carried out under an independent traffic pattern, as described in Section 5.2. As a set of parameters used in the simulation, the sizes of the internal DSE buffer and input buffer of the switch are set to zero. The dilation configuration is changed to [2,3,4,variables] and [2,4,6,variables], for which the term "variables" refers the dilation of the last stage.

Figure 34 shows the cell loss probability of a 16x16 FAB switch in a function of last stage output dilation for two different dilation configurations [2,3,4] and [2,4,6]. For dilation configuration [2,3,4,variables], the minimum cell loss probability is in the order of 10^{-3}. The cell loss probability stays flat as soon as the last stage of dilation goes beyond 5. For the dilation configuration [2,4,6,variables], the cell loss probability is in the order of 10^{-7}, and starts to saturate when the last stage of dilation is greater than 8. This implies that the cell loss probability does not have a significant improvement as the dilation keeps increasing.

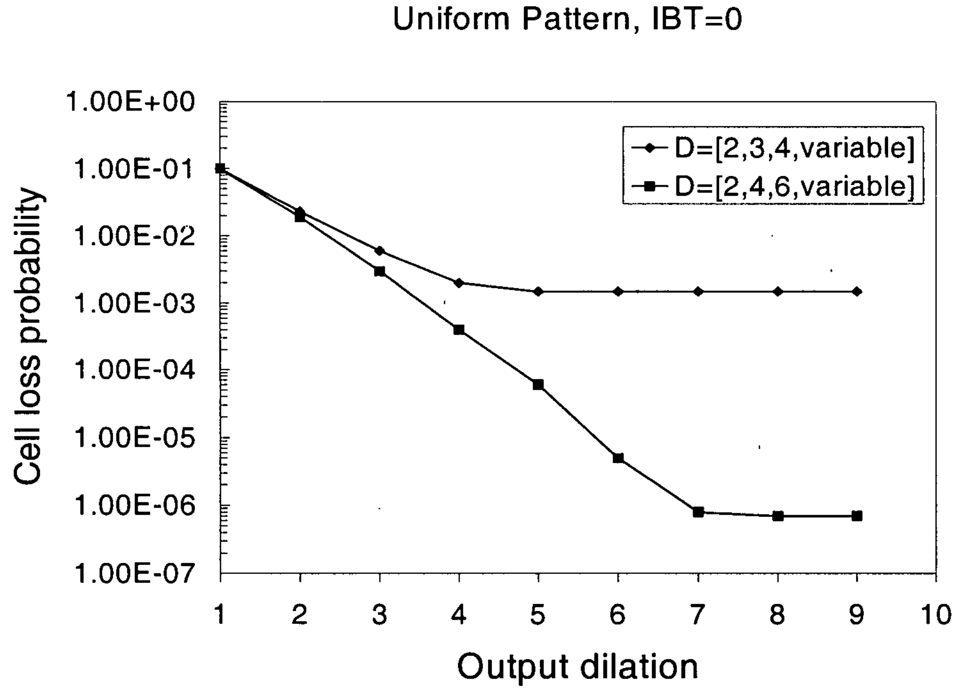


Figure 34 : 16x16 Banyan Switch under Uniform Traffic with IBT =0, OBT=60 and BS =0

Although the performance of the ATM switch core under a bursty traffic pattern, and the effect of the back-pressuring, are the main issues to be explored, additional simulations are performed to justify the robustness of the switch. All simulation results will be categorized and analyzed into several sections. Section 5.5 shows the performance of the advanced ATM switch core with a sliding window scheme and back-pressure control. Section 5.6 shows the performance of the ATM switch without a sliding window, while the back-pressure control scheme still exists. Section 5.7 shows the performance of the ATM switch with no sliding window mechanism or back-pressuring. Section 5.8 shows the results under a multicast bursty traffic pattern under all the above mentioned features. All simulation results maintain at 95% confidence interval.

5.5 With Sliding Window and Back-Pressure

Comparisons are made between Figure 33 of this thesis and Figure 9 of [16]. For the two different dilation configurations, the proposed model in this thesis under uniformly distributed traffic, can achieve one order of improvement in cell loss probability, as compared to a similar FAB switch (see Figure 9 on [16]).

First, the effect of output buffer size on an output-buffered switch is studied. An output buffering switch can achieve optimal throughput and delay performance, but a large buffer is needed to accomplish these results. Figure 33 is a plot of cell loss probability as a function of output buffer size per port on the dilation configuration [2,4,8,8], for the internal switch buffer (BS) from 0 to 5, under an independent uniform traffic pattern. For the dilation configuration [2,4,8,8], the cell loss probability curves decrease by one order of magnitude when every two internal buffers (BS) are added. This indicates that the addition of an internal switch buffer lowers the cell loss probability on an FAB switch core. In a later section, performance improvement due to the increase of the internal buffer size (BS) under bursty traffic will be shown. If BS is 0, and the output buffer per port is 60, the cell loss probability drops to less than $6*10^{-10}$. This result, when compared to that presented in Figure 10 of [16], has shown an improvement of one order in terms of cell loss probability. It is important to note that the FAB switch proposed in this thesis, which has a back-pressure mechanism, achieves a comparable performance to the Knockout Switch [7]. Nevertheless, a Banyan structure requires a much lower hardware complexity than a Knockout structure.

For a bursty traffic condition, a larger input, output, and internal FSE buffer size per port are needed for handling large successive incoming cells from the same destination address. For the internal switch core that is at a high fixed load (i.e. 0.9), congestion will occur when the output dilation is not fully dilated, (i.e. dilation configuration [2,4,8,16]) or it will reach the threshold of the internal buffer or input (output) buffer. Figure 35, 36 and 37 show the simulation results for a 16x16 FAB switch under different dilation configurations and 3 different sets of total buffer size per port (300, 500 and 1000). The results indicate that the throughput under burst traffic would increase as the output dilation of the SE increases, and hence, the cell loss probability decreases.

By varying the ratio of the input buffer to the output buffer, the results successfully locate the optimal choice of the input and output buffer for a specific cell loss probability. The optimum point is easily observed at the ratio of IBT/OBT being 0.9, when the sum of IBT and OBT is 300, 500, and 1000. When the total buffer size per port is set to 1000, cell loss probability becomes less than $6.9*10^{-8}$ for all the different ratios of IBT/OBT on dilation configurations [2,2,2,2] and [2,4,8,8].

Before comparing the results obtained in this thesis with the previous results (Figure 14 of [16]), both the internal buffer size and input buffer size are set to zero so that cell loss will occur. Figure 38 is a plot of cell loss probability as a function of output buffer per port using an average bursty length equal to 15. The proposed model achieves a better cell loss performance (see Figure 38) on all different output buffer size per port compared to figure 14 of [16]. For example, when the output buffer size per port is 500, the proposed model accomplishes a cell loss probability of $7.6*10^{-6}$, while the switch model from [16] obtains a cell loss probability of $2e^{-4}$. On the other hand, cell loss probability at buffer size 700 is less than $7*10^{-10}$, as shown in Figure 38, whereas the cell loss probability from [16] is $3*10^{-5}$ under the same set of parameters. Hence, there is a 2 order of magnitude of improvement when the output size per port is 500, and at least a 4 order of magnitude of improvement when the output buffer size per port is 700.

By increasing the internal buffer size to 10, a further improvement on the cell loss performance can be made. At least a 3 order of improvement is achieved (see Figure 36) when the output buffer size is 500. Notice that it has the dilation configuration [2,4,8,8] as opposed to the TBFT(16,3) D=[2,4,8] in Figure 14 of [16].

The following information is important in order to understand the behavior of the output dilation of the SE at different stages. All inputs, internal switches and output buffer sizes must be reduced so that cell loss probability higher than $6.9*10^{-8}$ can be observed (this is a limitation of the 1 million simulation cycles with 16 input ports and load at 0.9). Figure 39 shows the cell loss, with a smaller total buffer size per port (both input and output port buffer sizes are set to 100) and load (p) condition varied between 0.7 and 1. Under these parameters, as the output dilation is greater than 2, the improvement in performance becomes negligible at a high load. Although the dilation does enhance the internal bandwidth, the output rate is the same as the input rate so that the contention problem will occur at the output buffer under hot spot traffic or multi-cast traffic. The implementation of a back-pressure mechanism could reduce the chance of cell loss at the output buffer. As it will be shown in Section 5.6, back-pressure signaling plays a major role on the output dilation configuration in reducing the cell loss probability. In Figures 35 to 37, the total buffer size per port increases to 300, 500 and 1000. Under each of these buffer sizes, the dilation configuration [2,4,8,8] always slightly outperforms the configuration [2,2,2,2]. Since the cell loss probability for both dilation configurations are more or less the same, the dilation configuration [2,2,2,2] is used for most of the simulations hereafter, unless it is otherwise stated.

Next, the effect of the internal switch buffer size on a bursty traffic pattern is studied. Figure 40 shows the cell loss probability when the load is at 0.9, input and output port buffer sizes at 150, and internal buffer size of SE varies from 0 to 20. The results reveal that cell loss probability decreases linearly as the internal buffer size increases. It contradicts the internal buffer behavior of the FAB without back-pressure, which is reported in [12, 13]. This strongly indicates that a further increased switch core performance can be achieved if the internal FSE buffer size is increased, as well.

To make the bursty traffic pattern more realistic, loads are varied under Poisson distribution over 30 million simulation cycles (see Figure 32). The FAB switch model is tested with an internal buffer size (BS) of 10. Results are shown in Figure 37. When the total buffer size per port is set to 1000, the cell loss probability is less than $2.6*10^{-9}$, which is the limitation of 30 million simulation cycles, 16 input ports and an average load at 0.8. The cell loss rate is slightly higher than the results under a high fixed load. The reason is that a high load at 0.9 repeats in succession; however, these results are acceptable.

A simulation is performed on a 64x64 FAB using the same set of parameters. Figure 41 shows a slightly worse result, as compared to Figure 35 and Figure 36. Cell loss probabilities on a 64x64 FAB for a 300 and 500 total buffer size per port are slightly higher than that in a 16x16 FAB. This is a worse result because of the presence of internal contention in a 64x64 under self-routing property. Moreover, the optimum point is slightly shifted to the left, and located at IBT/OBT equal to 0.3.

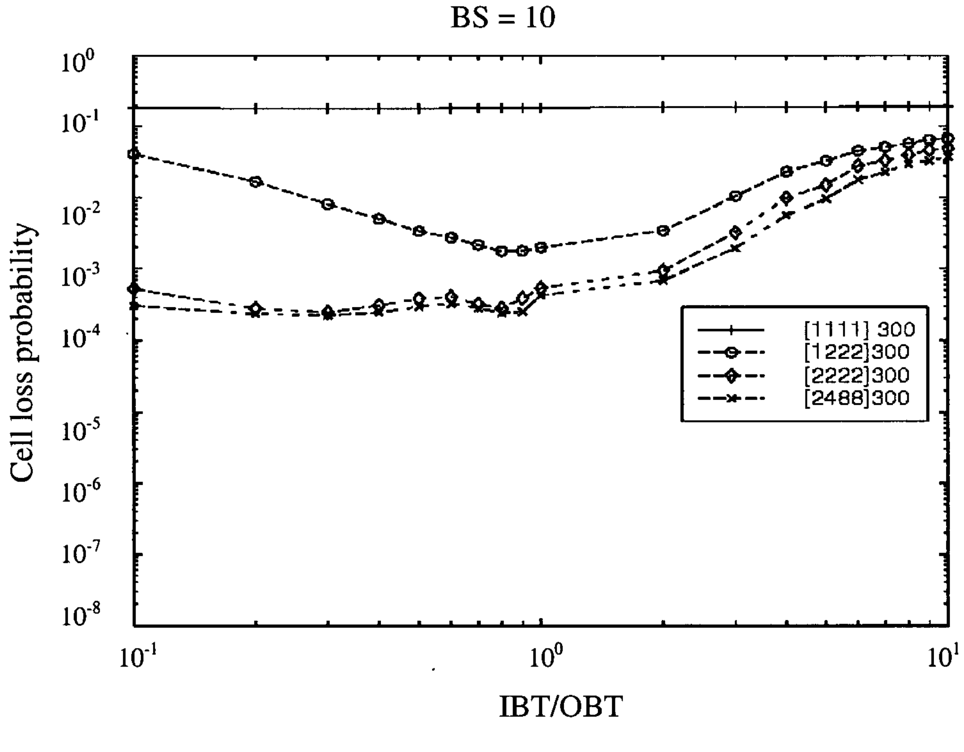


Figure 35 : 16x16 Banyan Switch Cell Loss vs 300 Total Buffer Size Per Port

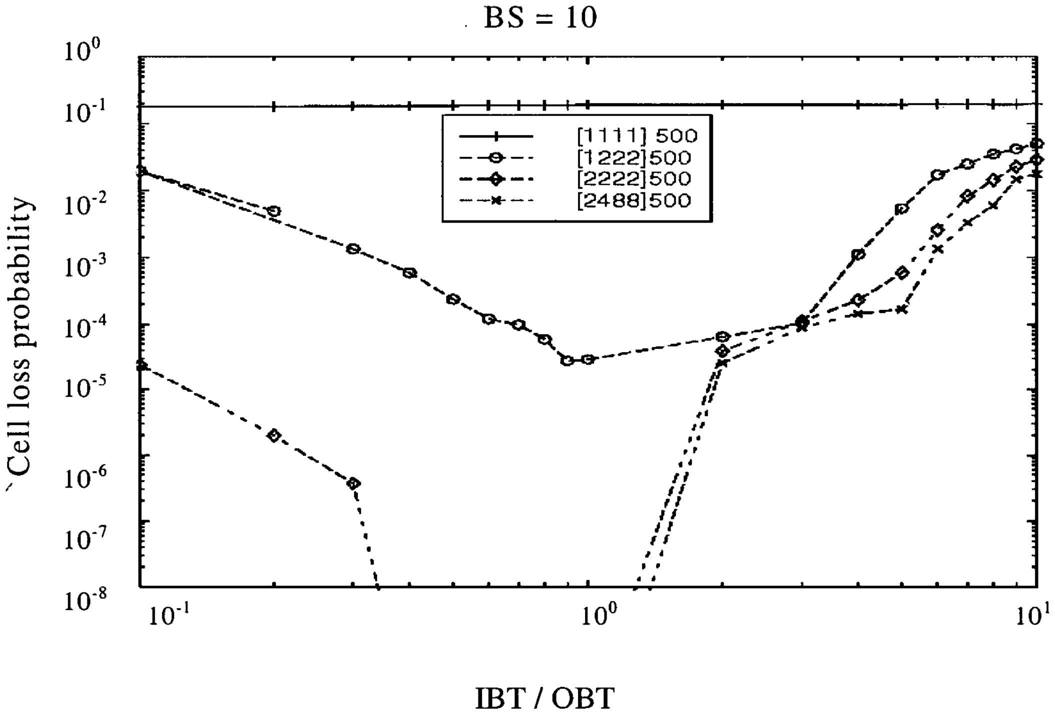


Figure 36 : 16x16 Banyan Switch Cell Loss vs 500 Total Buffer Size Per Port

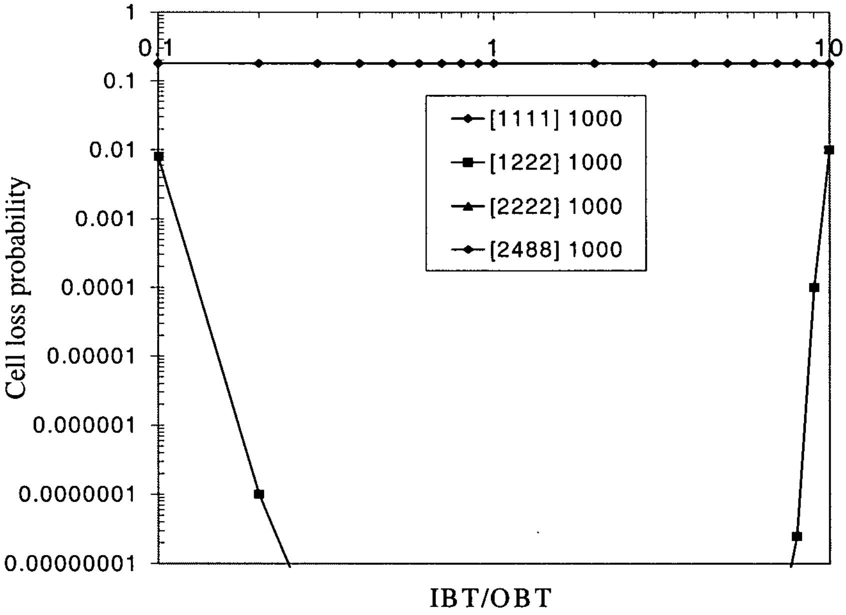


Figure 37 : 16x16 Banyan Switch Cell Loss vs 1000 Total Buffer Size Per Port

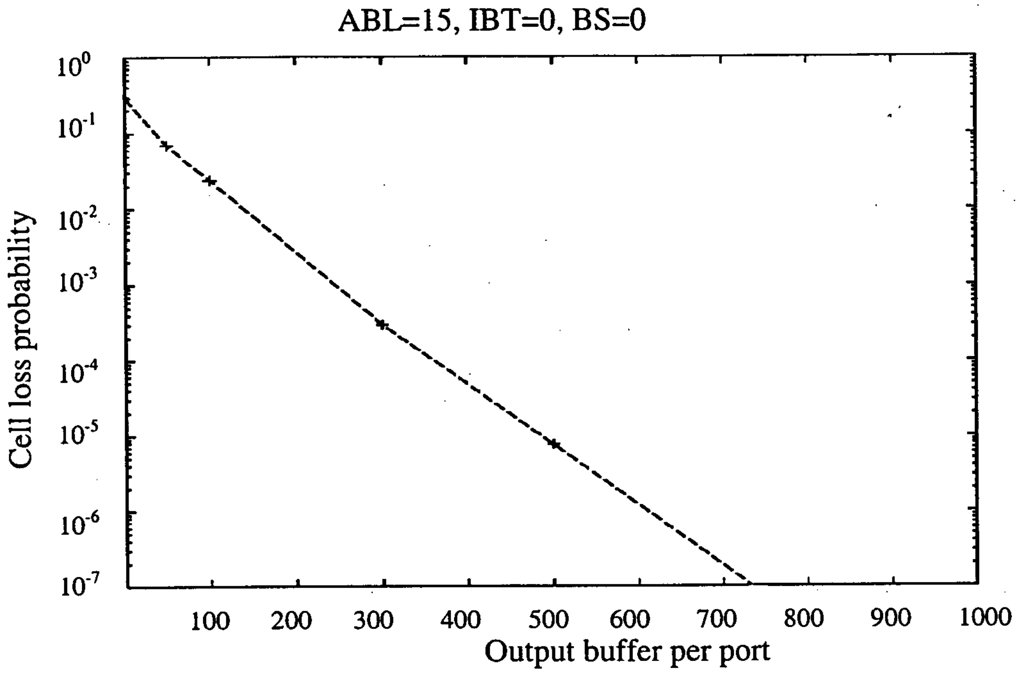


Figure 38 : 16x16 Banyan Switch with no Input and Internal Switch Buffer under Unicast Bursty Traffic Pattern

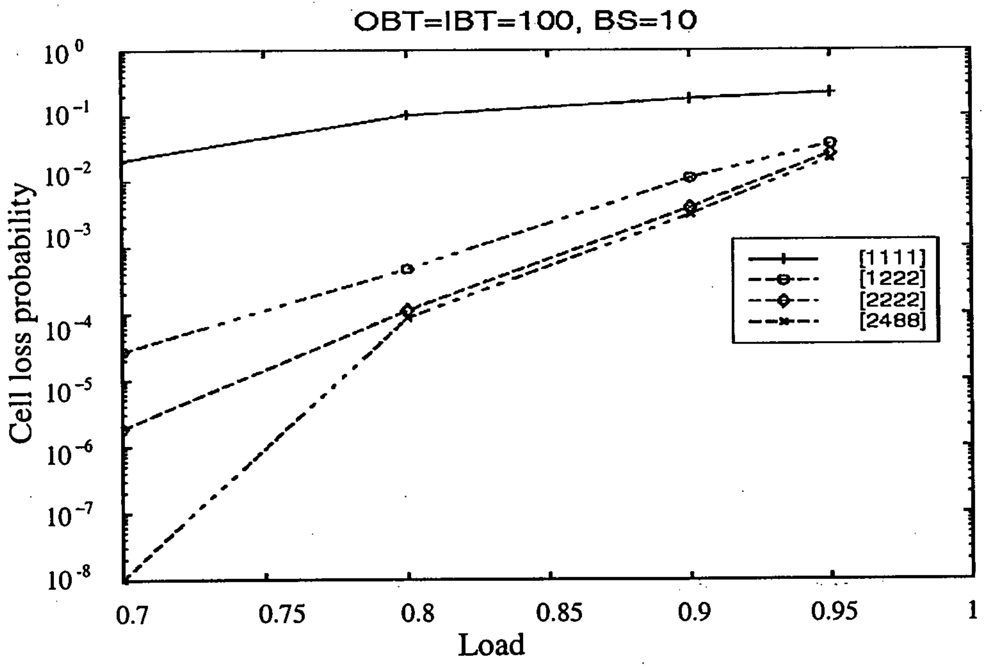


Figure 39 : 16x16 Banyan Switch Cell Loss vs Load

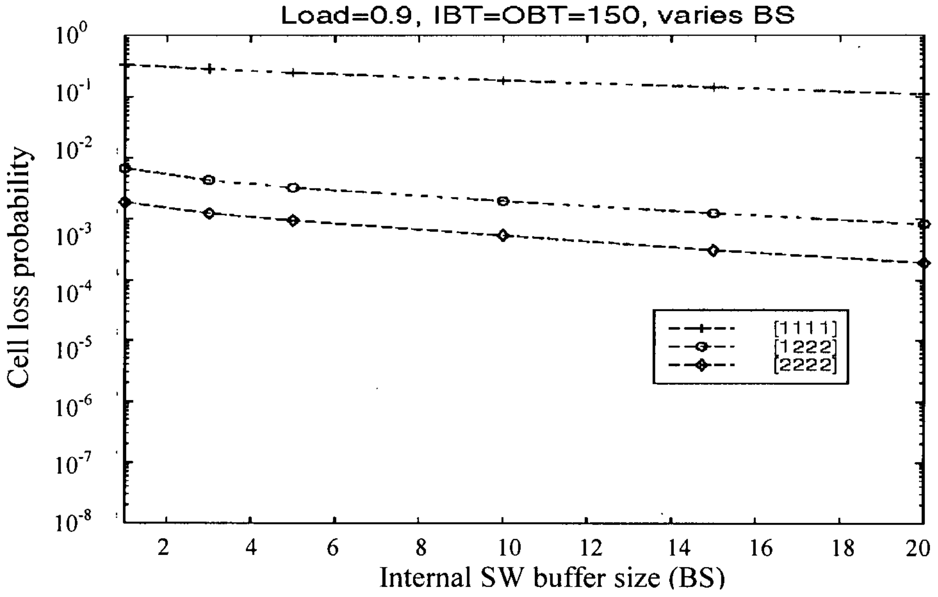


Figure 40 : 16x16 Banyan Switch Cell Loss vs Internal Switch Buffer

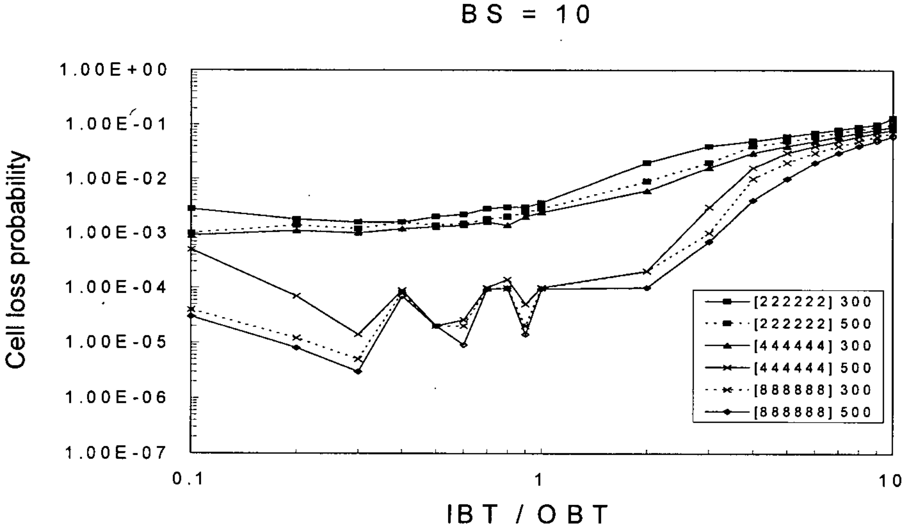


Figure 41 : 64x64 Banyan Switch Cell Loss vs Total Buffer Size Per Port under Multicast Bursty Traffic

5.6 With Backpressure Control but no Sliding Window

The performance of the proposed 16x16 ATM switch cores, without a sliding window while keeping the back-pressure control, is reported as follows. The same set of network parameters (average bursty length is 15, total buffer sizes per port are 300,500, & 1000, and internal FSE buffer size is 10), and the same dilation configuration [2,2,2,2] that was previously used in Figure 26, is used in this simulation. Low performance is expected as it was stated in Section 4.1 that the sliding window mechanism can remove the HOL blocking effects on an internal switch buffer, and input and output port queues. Figure 42 is compared with the curves (dilation configuration [2,2,2,2]) in Figure 35, 36, and 37. Without sliding window features, a high cell loss probability is reported. The sliding window mechanism does provide improvement of at least one order of magnitude in the cell loss performance at the optimum point (IBT/OBT = 0.7).

Figure 43 illustrates the results of the various SE internal buffer sizes (BS) at a high fixed load 0.9. The diagram clearly shows that there is no significant improvement as the BS increases without a sliding window

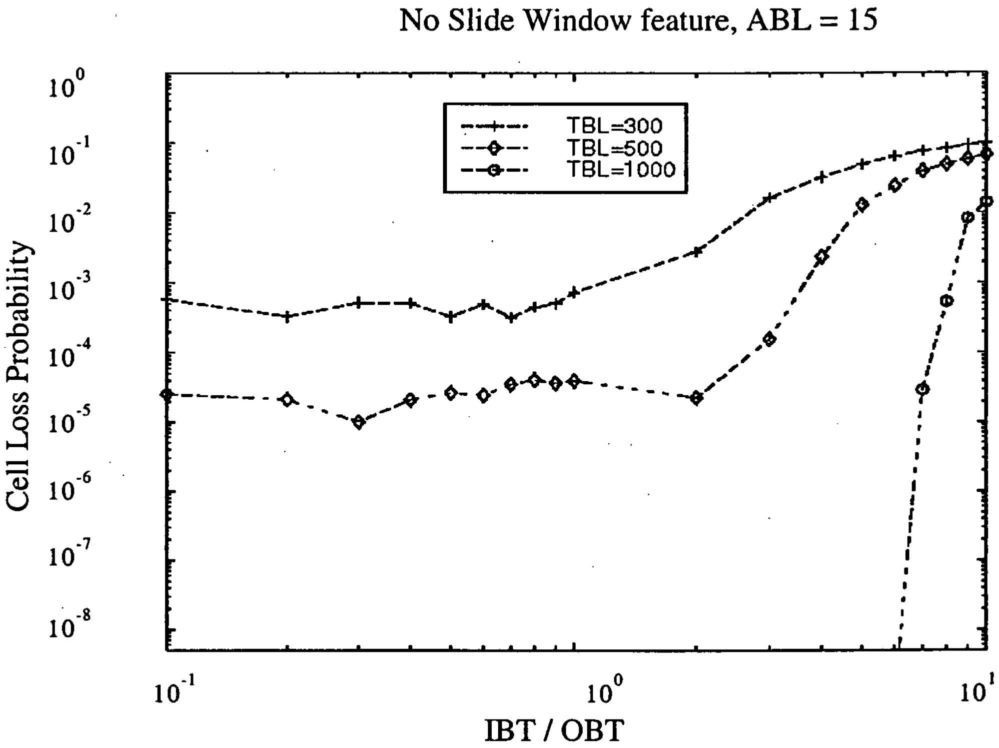


Figure 42 : 16x16 Banyan Switch Cell Loss vs IBT/OBT Ratio without Sliding Window Feature

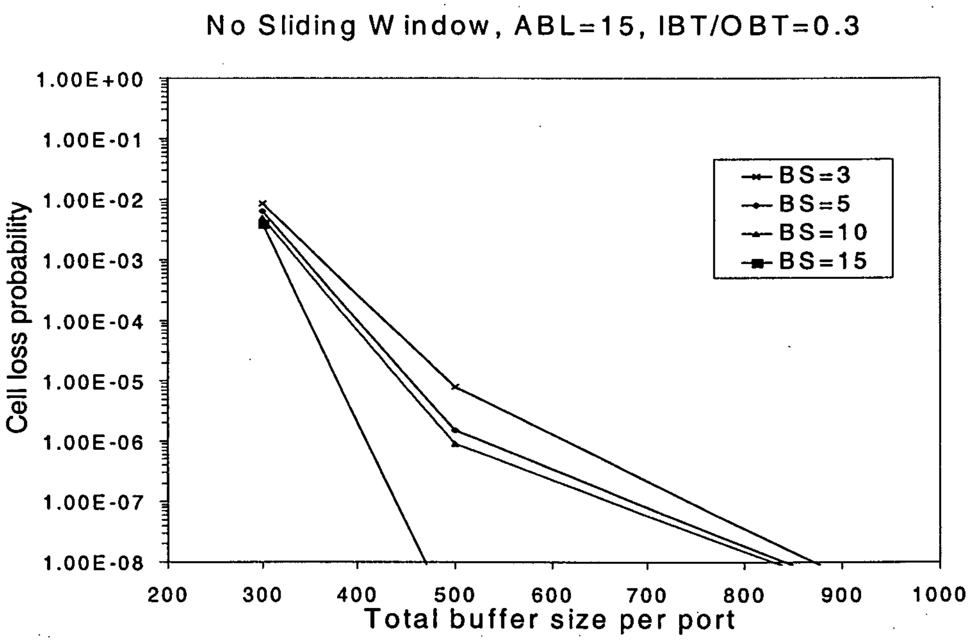


Figure 43 : 16x16 Banyan Switch Cell Loss vs Total Buffer Size Per Port on Different BS without Sliding Window Feature

5.7 Without Sliding Window and Back-Pressure

Without any additional features, the proposed ATM model becomes the first generation ATM switch core. The cell loss probability is extremely high, approximated at 0.01 for all 3 sets of total buffer size per port. These results are not acceptable. The sizes of input and output port buffer are insignificant, as well as the internal FSE buffer size. The cause of high cell loss probability is due to internal congestion for the dilation configuration [2,2,2,2]. For the fully dilated configuration [2,4,8,16], losses will occur at the output port buffer. Since back-pressure is not present, the internal SE buffer and input port buffer will not play a role in avoiding the cell loss. A low cell loss probability can still be achieved if an extremely large output buffer port is used to hold the bursty cell. Due to the high cost, a large output buffer port will not be considered.

5.8 Results for ATM Pattern

Figures 44 and 45 below show low cell loss performance results under a multicast bursty traffic pattern with an average burst length of 15, and Fanout equal to 2. At a 0.9 output effective load, cell loss probability is maintained below 10^{-9} as the internal switch buffer size (BS) increases to 25.

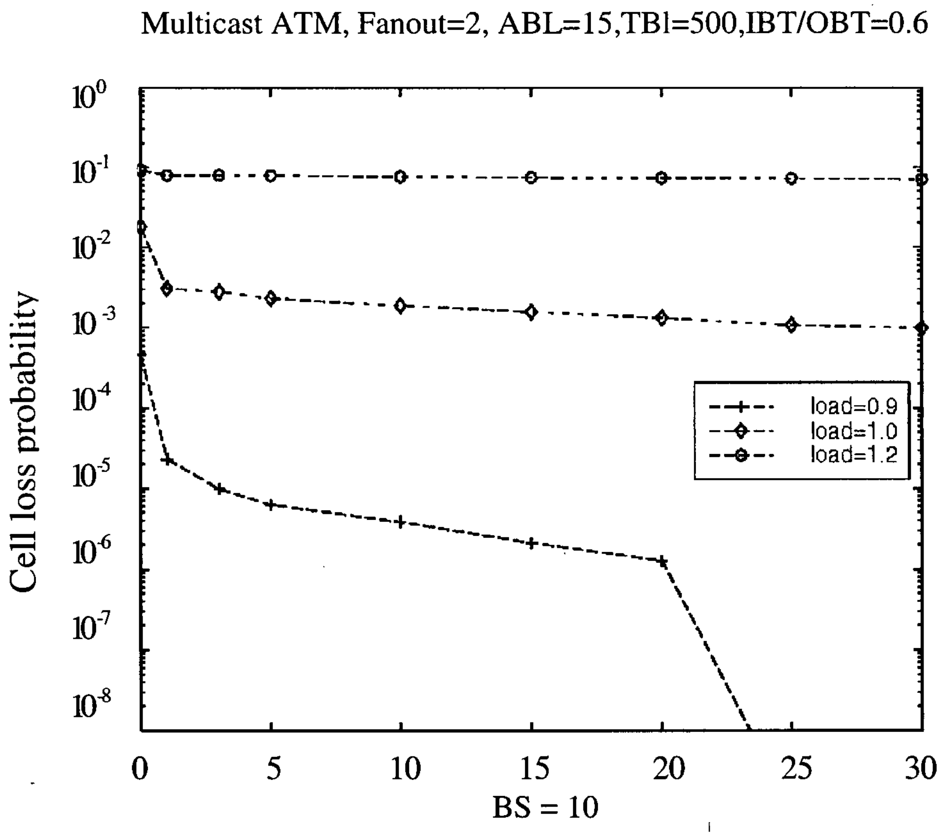


Figure 44 : 16x16 Banyan Switch Cell Loss vs Internal Switch Buffer (BS) with FANOUT=2 and under Multicast Bursty Traffic (ABL=15)

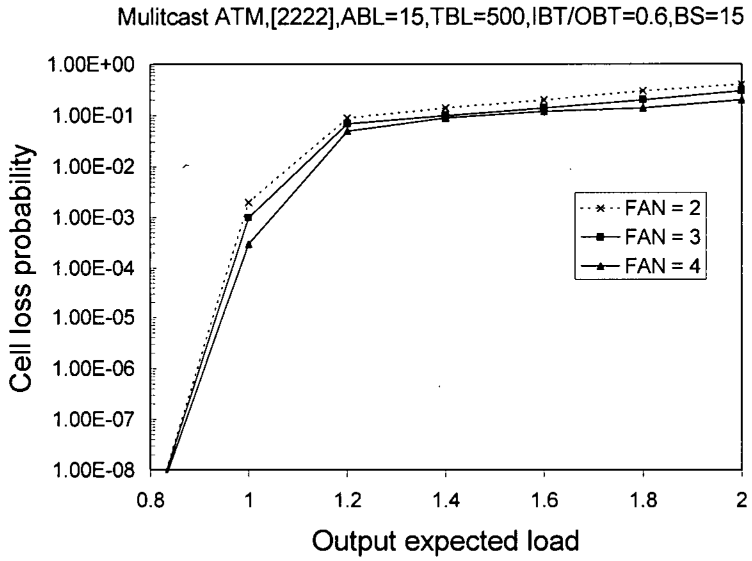


Figure 45 : 16x16 Banyan Switch Cell Loss vs Output Expected Load with FsANOUT=2 and under Multicast Bursty Traffic (ABL=15)

Figure 46 shows an optimum point on the allocation of the input and output buffer size (per port) under a bursty traffic pattern, with an average burst length of 15 and 0.9 load. As shown in Figure 46, a cell loss probability lower than 10^{-9} is achievable with the proposed switch even though it is at high loads and bursty traffic. For multicast traffic, the traffic model considered is a multicast connection, which constitutes a given fraction of the traffic, and has a fixed average fanout (i.e. number of copies per input multicast packet).

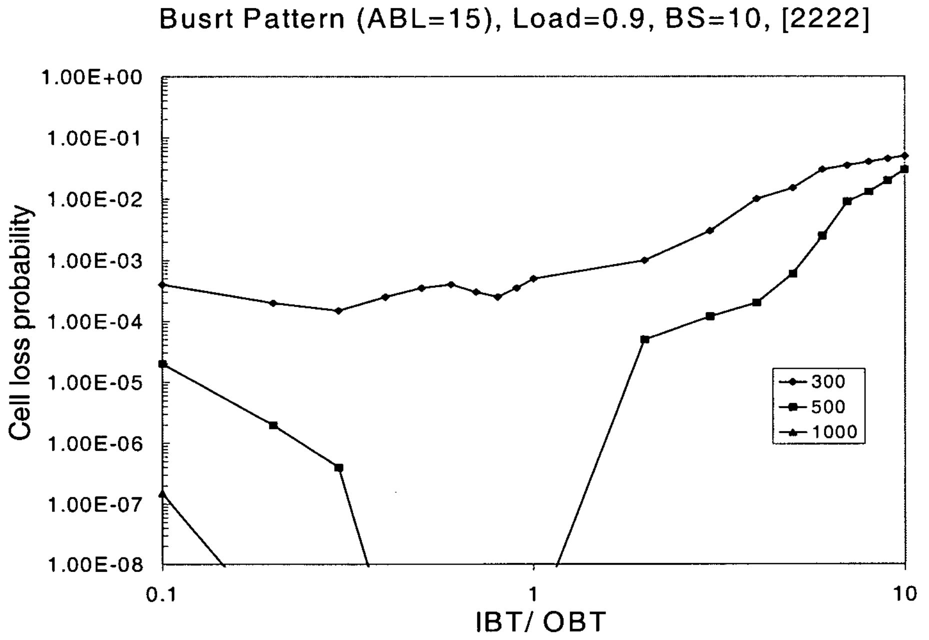


Figure 46 : Cell Loss vs Ratio of IBT/OBT under Bursty ATM Traffic

Figure 47 shows superior cell loss performance under a multicast bursty traffic pattern, with an average burst length of 15, and multicast fanout of 3. At 1.0 of effective output load, cell loss probability is maintained under 10^{-9} when the multicast-to-unicast cell ratio (M:U) is 0.1, 0.2, and 0.3, and the total buffer size per port is increased to 1000.

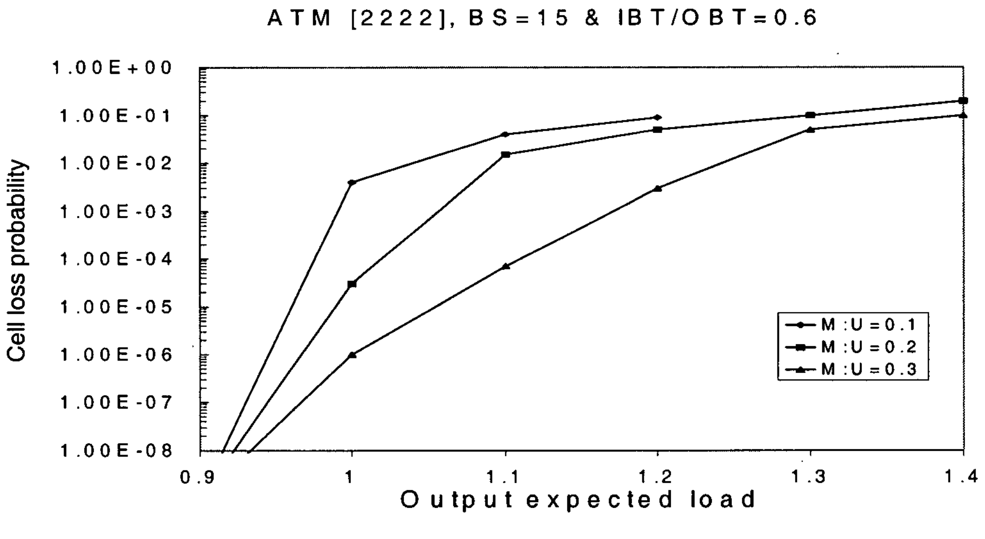


Figure 47 : Cell Loss vs Output Expected Load under Multicast and Unicast ATM Bursty Traffic without RED Meachanism

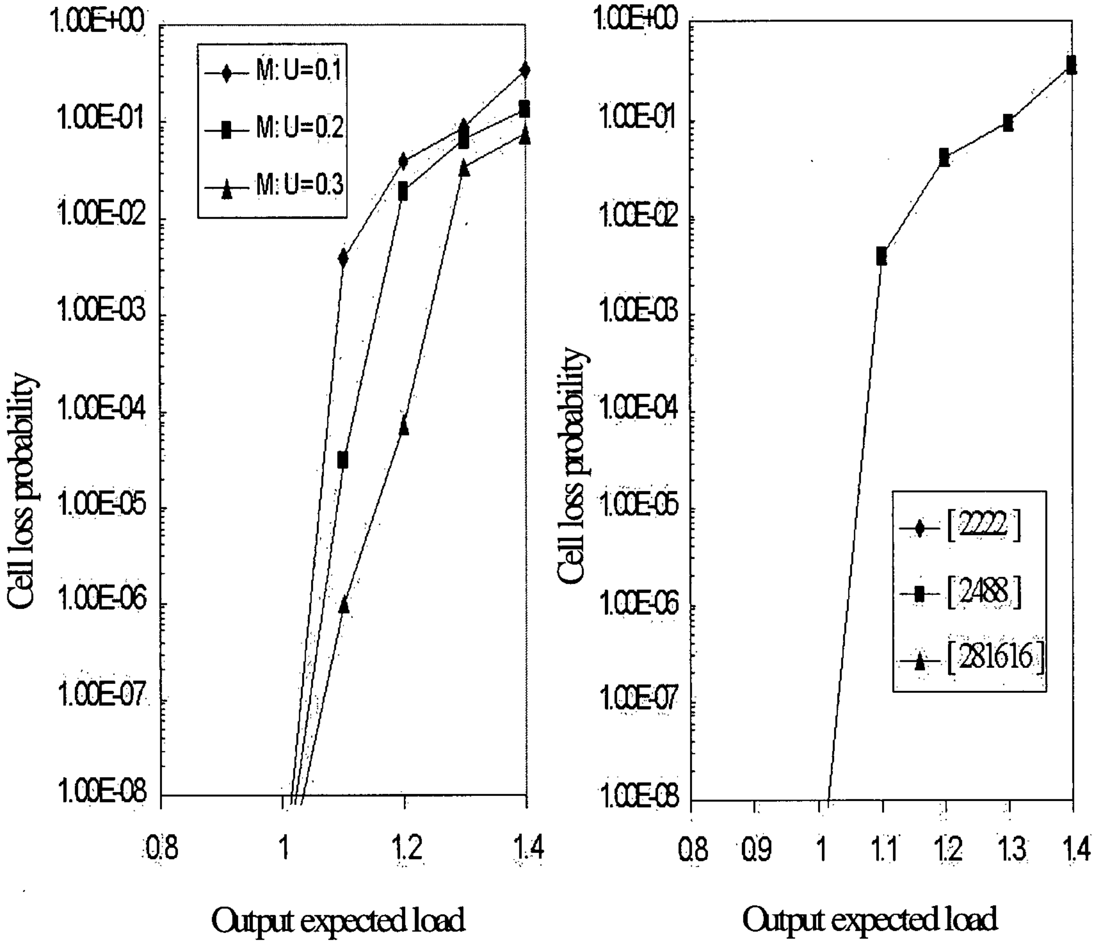


Figure 48 : ATM, ABL=15, Fanout=3, BS=15, IBT/OBT=0.6, IBT+OBT = 1000

Figure 48 shows the average queue delay on each expected output load traffic. The queuing delay for an ATM cell is determined by the time difference between when the ATM cell arrives at the input of the FIFO and the time that the same ATM leaves the FIFO. When the expected output load is 1.0, the queuing delay that is less than 20 can have an outcome of an M:U ratio equal to 0.1, 0.2, and 0.3.

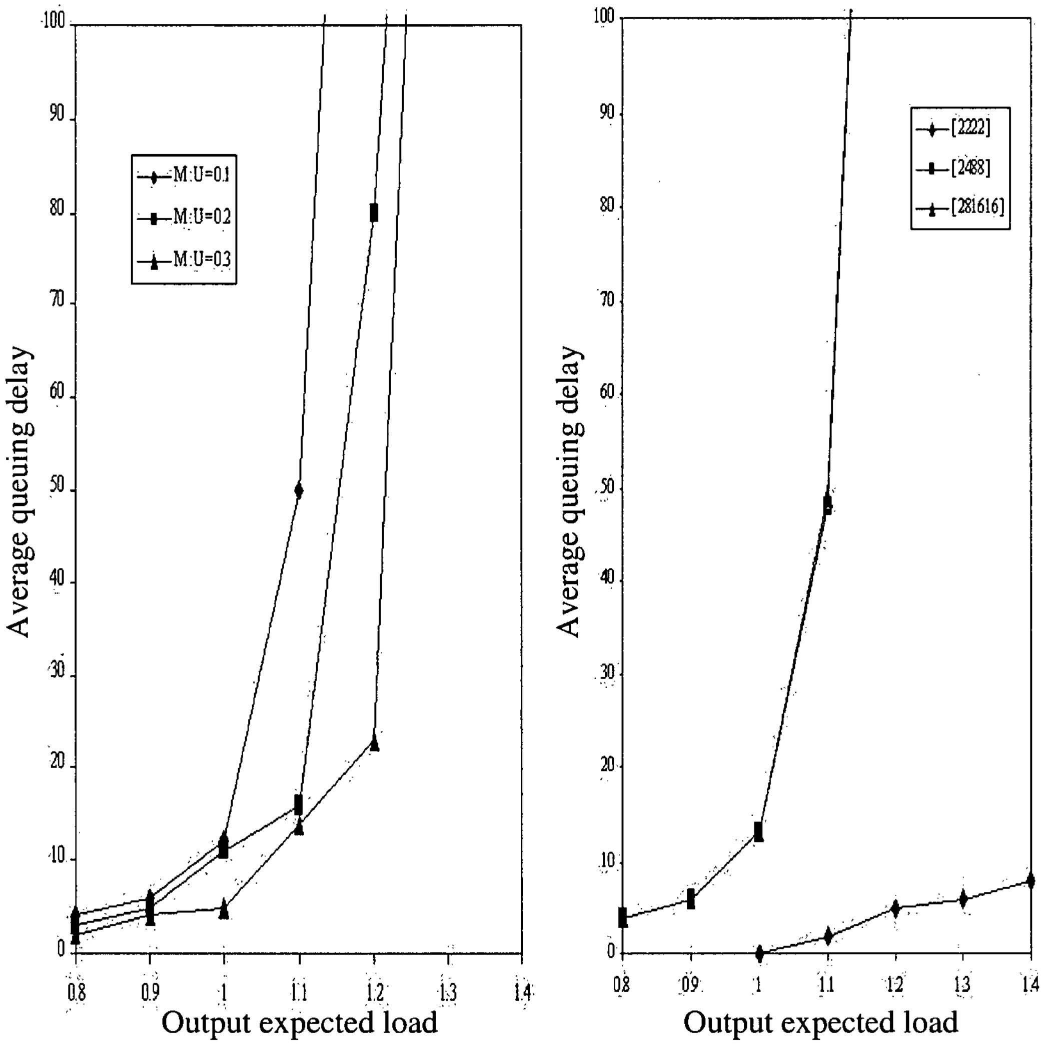


Figure 49 : ATM[2222], ABL=15, Fanout=3, BS=15, IBT/OBT=0.6, IBT+OBT = 1000

5.9 Results for IP Pattern

In the circumstance of the IP switched-router, an input traffic model has two priority classes: high-priority (H) and low priority (L). Figure 50 shows a very low packet loss performance under bursty input multicast traffic with different average burst lengths (10, 15, and 20) and a multicast fanout equal to 3. In all cases, a RED algorithm with two drops precedence was used. Figures 51, 52, and 53 illustrate the delay, and good output performances, for both low and high priority traffic at different loads, and average burst lengths. The higher priority class packet always achieves a better performance on both packet loss and routing delay.

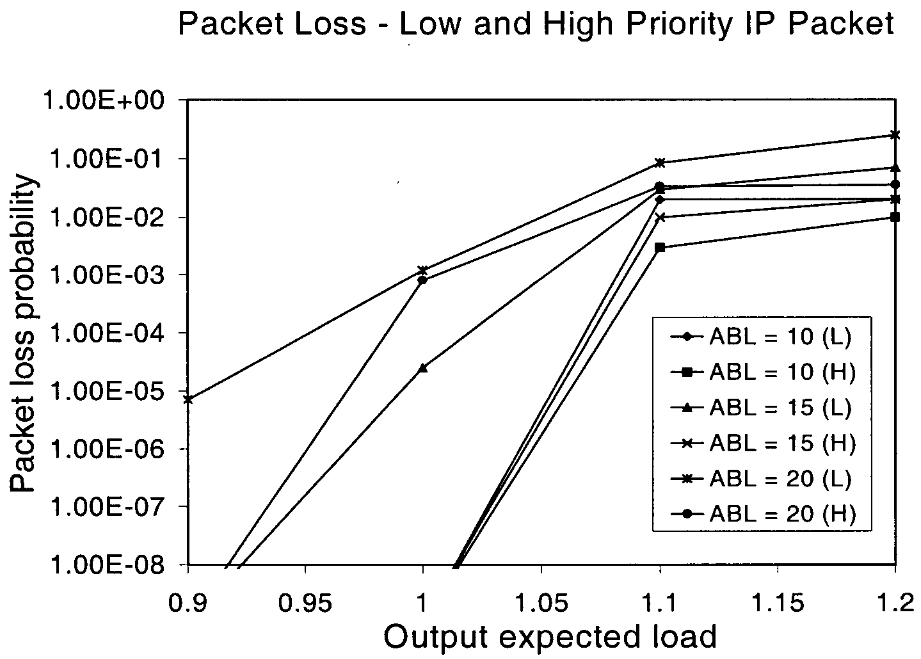


Figure 50 : 16x16 Ultra Switch Packet Loss vs Expected Output Load under IP Multicast Bursty Traffic with FANOUT=3 and RED Mechanism

Delay latency on an IP packet is determined between the start of packet's arrived time at the input of FIFO and the end of the same packet's leaving time at the output of FIFO. At the output expected load of 0.9, delay latency on a high priority IP packet is found to be double that of its average burst length, as shown in Figure 51. This implies that a minimum delay latency is achieved. This is because, for example, when the data of an average burst size of 10 is received at the input of FIFO, it takes 10 simulation cycles to receive the completed 10 packets. Since the output FIFO will not transmit the packet until the numerous completed data is received at the output FIFO, it takes another 10 simulation cycles to transmit the data at the average burst size of 10. Therefore, the minimum delay latency is the sum of 10 and 10, which doubles the average burst size. At an expected high output load, delay latency dramatically increases as the packet loss ratio increases due to contention.

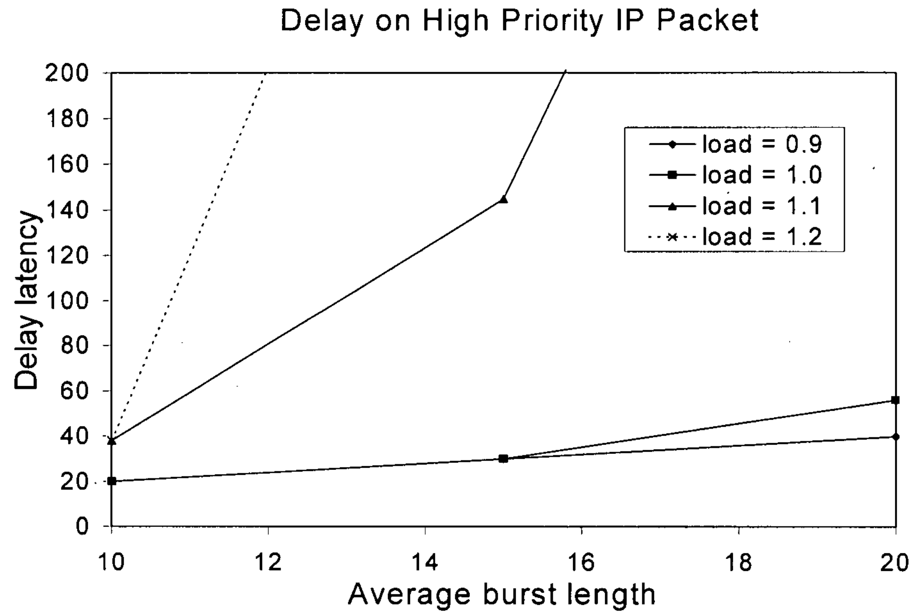


Figure 51 : Banyan Switch Latency on High Priority Class vs Average Burst Length under IP Multicast Bursty Traffic with FANOUT=3 and RED Mechanism

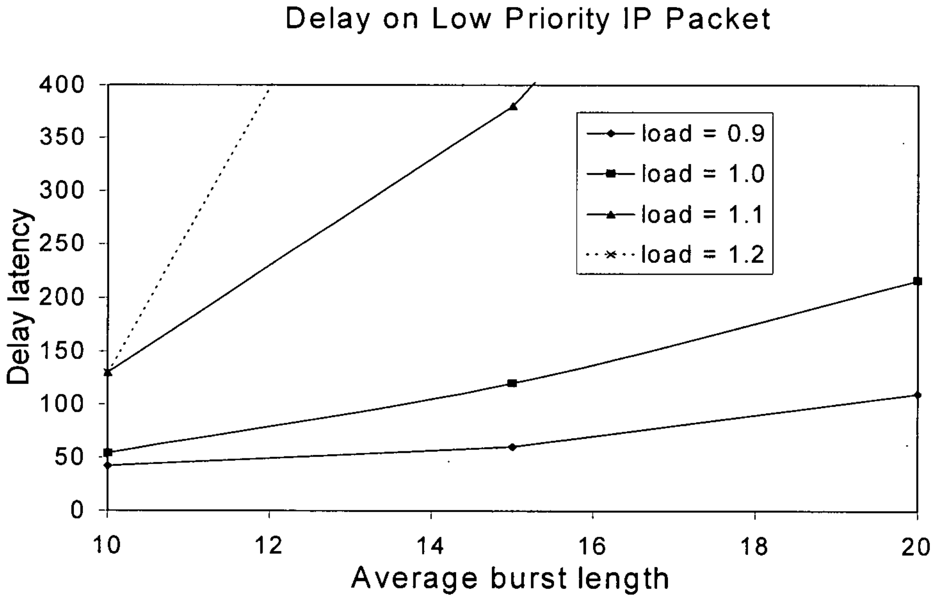


Figure 52 : 16x16 Banyan Switch Latency on Low Priority Class vs Average Burst Length with FANOUT=3, IP Multicast Burst Traffic and RED Mechanism

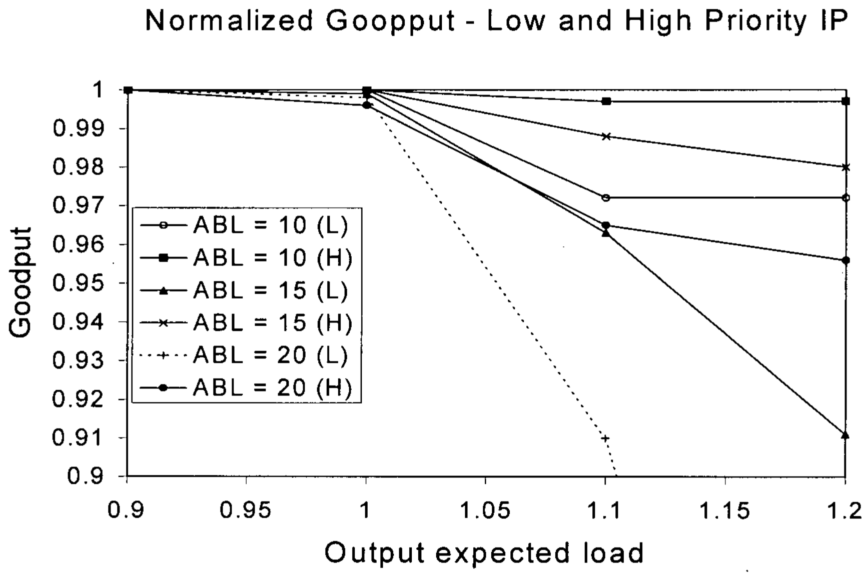


Figure 53 : Normalized Goodput vs Output Expected Load with Different Burst Length (10,15 and 20), FANOUT=3 and RED Mechanism.

To prove the RED mechanism does improve the average queue length, the same set of parameters are used. These are the dilation [2222] , internal buffer size of 15, and FANOUT =3. From Figure 50 and Figure 57, the packet loss ratio and delay latency of a high priority packet is applied on both with and without a RED mechanism. However, these do not show any significant difference. The delay latency of a low priority packet, that is under a RED condition, tests at a better result. At a 0.9 output load, the delay latency is expected to result in a low priority packet since no packet is lost. When the expected output load increases, the delay latency of a low priority packet with RED shows a significant improvement because RED starts to drop packets when the input FIFO has reached its 90% capacity. A large packet is more likely to be dropped than a small packet. Although RED starts dropping packets when the input FIFO reaches its RED threshold, the packet loss ratio with RED is not significantly greater than the outcome that is without RED.

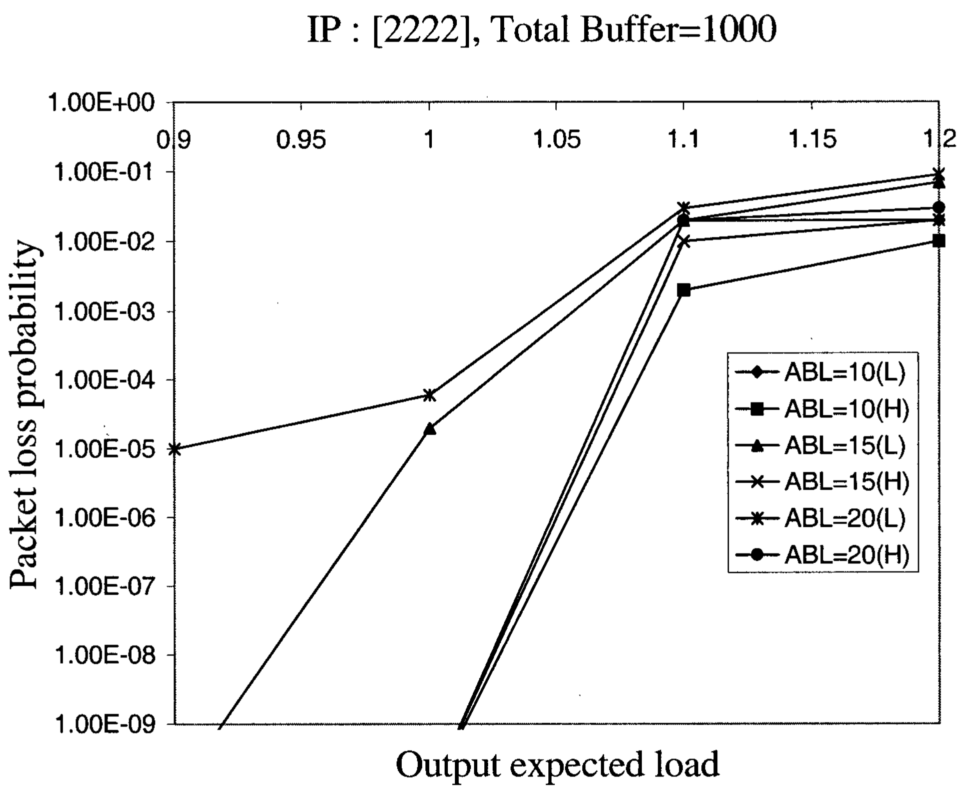


Figure 54 : IP traffic [2222], NO RED, BS = 15

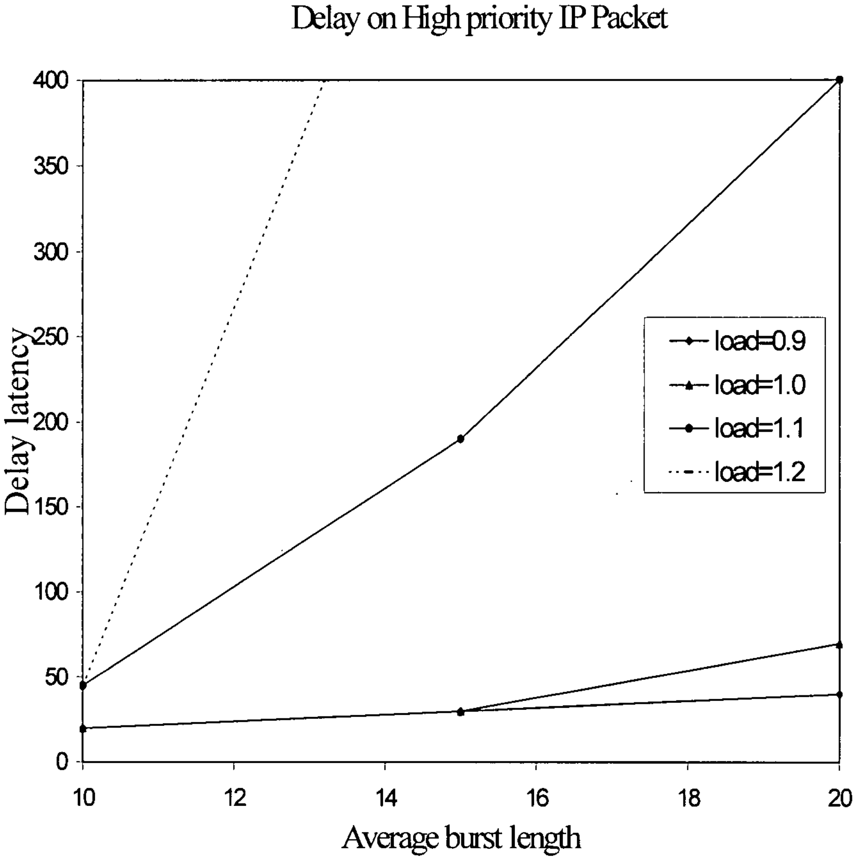


Figure 55 : Delay on a High Priority IP Packet, [2222], NO RED BS=15

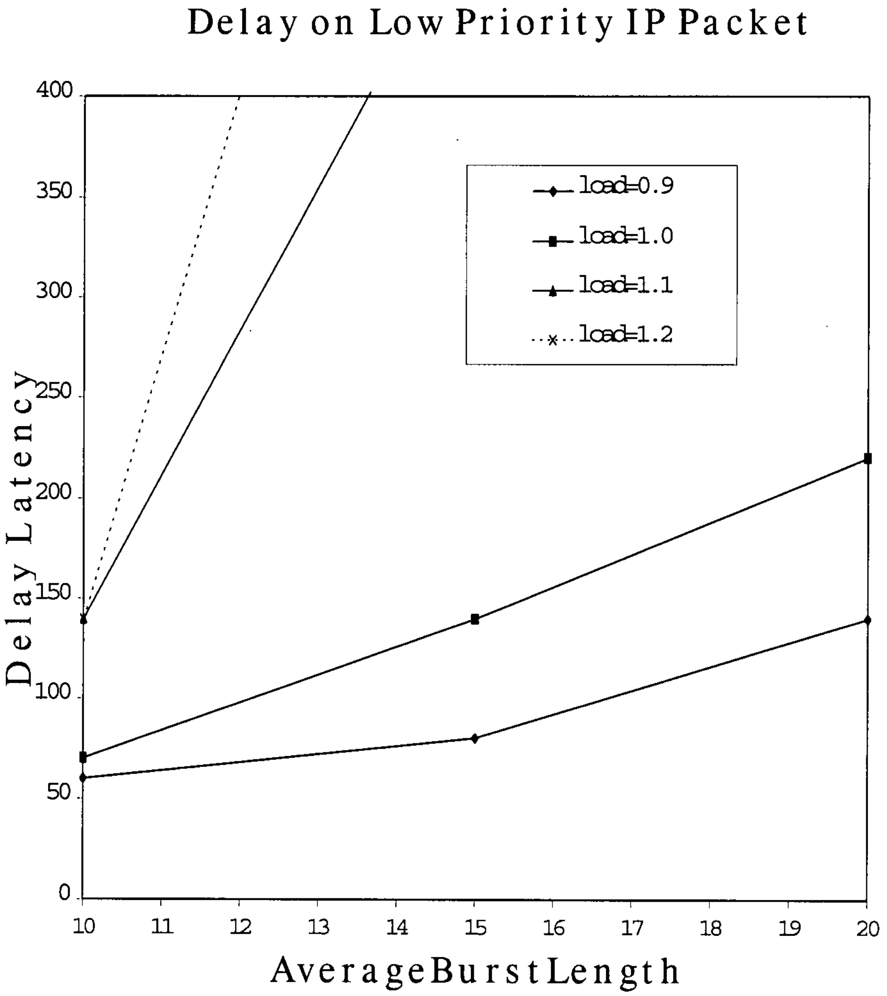


Figure 56 : Delay on a Low Priority IP Packet, [2222], NO RED, BS=15

Chapter 6 Conclusions and future research plan

6.1 Conclusion

In this thesis, a high performance ATM/IP switch that can achieve very high throughput and low cell/packet loss rate is presented. Simulation results are obtained under all of the major kinds of cell traffic patterns that most research is focused on, namely uniform, bursty, single-class, and multi-class on unicast and multicast traffic patterns. With high throughput and a low cell/packet loss rate, these results confidently imply that this switch is practical in today's networking industry.

The back-pressure mechanism and internal switch buffer are two major parameters related to the performance of the proposed switch. From the results of these simulations, back-pressure control shows an important improvement on cell loss probability, as does the sliding window feature. Back-pressure control will increase the complexity on the design. However, improvement of at least 3 orders on the cell loss probability can be achieved, as shown at the optimum operating point. With back-pressure control, the internal buffer becomes significant. Although the sliding window is difficult to implement and construct in hardware, the improvement is of at least two orders at the optimum point.

6.2 Future Research Plan

The work in this thesis can be extended to further enhance the design of the ATM/IP switches. As a continuation of the work in this thesis the following areas can be investigated in future research work.

1) The simulation of the two class traffic models in this thesis is based on a constant ratio between high priority and low priority traffic. Simulations with various ratios are worthwhile to considered so that more realistic results can be obtained.

2) Simulations of multicast traffic simulation in this thesis are assumed to have a constant proportion of multicast traffic. Simulations of various proportions should be considered in order to obtain more realistic results.

3) Future work can be focused on the hardware implementation of the sliding window based on the research projects Tbit/s physical layout of the multi-FIFO design.

4) More simulation results should be obtained on hot spot traffic to increase the characterization coverage of the switch.

Appendix : A List of standard acronyms

AAL	ATM adaptation layer
ABL	Average burst length
ABR	Available bit rate
ATM	Asynchronous transfer mode
BDB	Buffered Dilated Banyan
BISDN	Broadband integrated services digital network
BP	Back-pressure
CAC	Connection admission and control
CBR	Constant bit rate
DRED	Dynamic Random Early Detection
DSE	Dilated switch element
EOP	End of packet
FIFO	First in first out
GFC	Generic flow control
HEC	Header error control
IP	Internet protocol
LAN	Local-area-network
MCR	Minimum cell rate
NNI	Network-network interface
OC-n	Optical carrier signal
QoS	Quality of service
RED	Random Early Detection
SOP	Start of packet
UBR	Unspecified bit rate
VBR	Variable bit rate
VC	Virtual channel
VCI	Virtual circuit identifier
VPI	Virtual path identifier
WAN	Wan-area-network

Bibliography

[1] M. Laubach, "Classical IP and ARP over ATM," Internet RFC 1577, Jan 1944.

[2] Broadband Publishing Corp., The ATM Report, vol4, no5, Aug./Sept. 1996

[3] E. Guarene, P. Fasano and V. Vercellone, "IP and ATM Integration Perspectives", IEEE Comm.Mag. Jan 1998, pp 74-80

[4] H.Ahmadi,W.E.Denzel, C.A.Murphy and E.Port, "A High Performance Switch Fabric for Integrated Circuit PacketSwitching", Proc. INFOCOM 88, pp 9-18.

[5] T Theimer, E. Rathgeb, and M. Huber, "Performance Analysis of buffered banyan networks," IEEE Trans Commun. Vol. 39, pp269-277, Feb. 1991.

[6] L.R.Gokr, G.J.Lipovski, "Banyan Networks for Partitioning Multi-processing Systems", Proc. First Annual Computer Architecture Conference, pp21-28,Dec.1973.

[7] Y.Yen, M.Hluchyj,A.Acampora,"The knockout switch:a simple,modular architecture for high performance packet switching", IEEE. Commun. vol5.pp1274-1283 Oct 1987

[8] E.T. Bushnell and J.S. Meditch, "Dilated multistage interconnection networks for fast packet switching," IEEE INFOCOM, pp1264-1273, 1991.

[9] M.Alimuddin, H.M. Alnuweiri, and R.W. donaldson,"The Fat Banyan,ATM switch," IEEE p. 659-666 INFOCOM, 1995.

[10] T.T. Lee and S.C. Liew,"Broadband packet switches based on dilated interconnection networks, "IEEE Transactions on Communications, vol.42, pp 732-744, Feb. 1994.

[11] M.Alimuddin, H.M. Alnuweiri, and R.W. donaldson,"Performance of the FAT-BAnyan switch under non-uniform traffic", IEEE p. 82-85 INFOCOM, 1995.

[12] Y. Mum and H.Y. Youn, "Performance analysis of finite buffered multi-stage interconnection networks," IEEE Trans. Commun., vol. 39, pp 269- 277, Feb. 1991.

[13] H. Yoon, K.Y. Lee, and M.T. Liu, "Performance of packet-switched banyans with arbitary switch sizes, queue sizes, lin kmultiplicities and speedups," IEEE INFOCOM, pp. 960-971, 1989.

[14] ATM switch Element User's Manual "WAC-188A", IgT co. June, 1997.

[15] R. Jain and S. Routhier, ``packet Trains - Measurement and a new model for computer network trafic," IEEE Journal of Selected Areas in Communications, vol.SAC-4,No.6, September 1986, pp. 986-995. Reprinted in Amit Bhargava, Ed., "Integrated Broadband Networks" Artech House, Norwood, MA, 1990.

[16] M.Alimuddin and H.M. Alnuweiri, "Design and Evalua ion of Scalable Shared-Memory ATM Switches", IEICE Trans. Commun, vol E81, pp224- 236,Feb 1998

[17] Sanjeev Kumar and Dharma P. Agrawal, "On Multicast Support for Shared-Memory-Based ATM Swithc Architecutre", IEEE Network. January/February 1996, pp 34-39.

[18] Fabrizio Sestini, "recursive Copy Generation for Multicast ATM Switching", IEEE/ACM trans. on Netwoking, vol 5, no 3, June 1997

[19] E.T. Bushnell and J.S. Meditch, "Dilated multistage interconnection networks for fast packet switching," IEEE INFOCOM, pp 1264-1273, 1991.

[20] Y. Mum and H.Y. Youn, "Performance analysis of finite buffered multi-stage interconnection networks," IEEE Trans. Commun., vol. 39, pp 269- 277, Feb. 1991.

[21] H.Ahmadi, W.E.Denzel, C.A.Murphy and E.Port, "A High Performance Switch Fabric for Integrated Circuit Packet Switching", Proc. INFOCOM 88, pp 9-18.

[22] Martin De Prycker, "Asyncronous Transfer Mode Solution for BroadBand ISDN", Prentice Hall, 1999.

[23] Dong Lin and Robert Morris, "dynamics of Random Early Detecction", SIG COMM'97 Cannes, Frances, pp 127-135.

[24] Floyd, S. Jacobson V., "Random Early Detection for Congestion Avoidance", IEEE/ATM Transactions on Networking, August 1993.

[25] C.M. Chu, H. Tayyar, and H.M. alnuweiri, " Enhanced Packet Switching on a Buffered dilated Banyan Switch", Proc of ISCA 14th International Conference on Computers and Their Application and Their Applications, Cancun, Mexico, pp 130-133. April 7-9, 1999

[26] T Theimer, E. Rathgeb, and M. Huber, "Performance Analysis of Buffered Banyan Networks," IEEE Trans Commun., vol. 39, pp 269-277, Feb. 1991.