Architecture of Reconfigurable a Low Power Gigabit ATM Switch

Abhijit M. Lele and S.K. Nandy
Supercomputer Education and Research Center
Indian Institute of Science
Bangalore 560 012 INDIA
[abhijit@,nandy@]serc.iisc.ernet.in

Abstract

Multistage switch interconnects like banyan switches are preferred in high speed networks for their cascadable structure and suitability for VLSI implementation. However most of these switch implementations are monolithic in nature and do not provide flexibility of dynamic re-routing of cells from active ports through idle ports. In this paper we take a critical look at a basic 8 × 8 *benes switch from the perspective of identifying smaller blocks which can be pipelined in space and temporally multiplexed to exploit hardware reuse. A topological analysis of a* 8 × 8 *benes switch is carried out to identify mutually exclusive path sets that can be overlayed for hardware reuse. Based on this analysis we arrive at a basic building block called* X-Structure, *using which a* 8 × 8 *switch is constructed. The* X-structure *supports dynamic re-routing of cells and power down mode. A communication controller is designed using the the* X-Structure *based ATM switch at its core. A performance evaluation of the switch indicates a power saving of 66.66% due to hardware reuse, an 18.6% increase in hardware utilization and an aggregate throughput of 2.66 Gbps for a* 8 × 8 *switch.*

1. Introduction

Switching systems are central to communication networks and forms the core element in any communication network. *Asynchronous Transfer Mode (ATM)* [1] networks have introduced a range of issues in ATM switch design. The two key features which guide ATM switch design are – the first being support for multi-rate traffic streams such as *Variable Bit Rate (VBR)* traffic with statistical multiplexing for efficient transmission of multimedia data and the second being support for multicast which enables switching of packets from one source to multiple destinations. ATM switches must address the issue of scalability, reliability and cost effectiveness. With recent advances in mobile computing, power efficiency needs to be addressed in the design of ATM switches as well.

A good survey of ATM switch architectures of the 90s can be found in [2] [3]. A more recent survey of ATM switch architectures appear in [4]. Some of the recent switch architectures are *Quality of Service Capable Switch* [5] and ATLAS [6] which uses shared memory architecture to implement the switch and the *Abacus* switch architecture [7] which is a scalable multicast ATM switch. The *Knockout Switch* [8] uses shared medium architecture for switching. The Helix-Switch [9] is a variation of the shared memory switch. Though efficient VLSI implementation are provided for all the proposed switch architectures, the inherent property of redundancy and possibility of hardware reuse is not well exploited to increase the switch throughput and efficiency.

In this paper we propose a novel switch architecture that seamlessly incorporates power optimization at all levels of the switch design *viz.* switching algorithm, switch architecture, and design of the switch element. The switch redistribute load from highly loaded ports to idle ports to achieve high level of hardware utilization. This effectively increases the switch throughput and helps reduce the switching latency.

The rest of the paper is organized as follows. In Section 2 we carry out a preliminary analysis of data traffic at input ports of ATM switch to motivate the fact that not all input ports of the switch are active all the time. We also carry out the topological analysis of a 8 × 8 switch to identify mutually exclusive paths. The design of an 8 × 8 switch is discussed in Section 3. The design of an ATM communication controller based on the proposed ATM switch is also discussed in this section. A performance evaluation of the proposed ATM switch is reported in Section 4, and we conclude in Section 5.

0-7695-0831-6/00 $10.00 © 2000 IEEE

2. Preliminary Analysis

Non-utilization of idle switch ports during high activity periods in conventional ATM switches leads to low hardware utilization resulting in low switching latency. In this section we carry out a preliminary analysis of a switching operation to identify sources of power optimization and justify the need for dynamic re-routing which in turn exploits hardware reuse.

2.1. Traffic Analysis

Simulations are carried out on a hypothetical 26 node network which is representative of a typical ATM network [13]. Each node in the network is an 8 × 8 ATM switch and the links in the network terminate at the ATM switch. For the proposes of simulation we assume there are no hot-spots and the destination addresses are uniformly distributed among all nodes and the duration of any session is uniformly distributed in the range 1 to 60 seconds. Simulations indicate that on an average, ports remain idle 28% of the time, and hence power saving to this extent can be achieved if the switching elements associated with idle ports are powered down to standby mode.

2.2. Hardware Reusability Analysis

Most of the multimedia traffic is bursty in nature and hence there are idle periods between two consecutive bursts of data at the input port of the switch. During this idle period the hardware associated with the corresponding input ports of the switch can be re-used to switch packets from other active input ports. The Hardware re-usability analysis is carried out to determine the extent to which the hardware can be re-used. Simulations indicate that the probability of idle slots across input ports tends to 0.186 for a traffic load of varying from 0.7 to 0.9. It may therefore be advantageous to utilize these idle slots to dynamically re-route packets from other active ports out of turn thereby reusing hardware 18.6% of the time.

2.3. Topological Analysis

A topological analysis of a 8 × 8 switch shown in Figure 1 indicates that using SW5, SW6, SW7, and SW8 as intermediate switches, connections are established between input and output ports as shown in Table 1. We define *Mutually Exclusive* paths to be the set of input-output pairs of ports that can be switched through one intermediate switch. A set of mutually exclusive paths forms a *Switching State*. An intermediate switch is associated with every *Switching State* as given in Table 1. A maximal set of mutually exclusive paths is defined to be a *Supper Channel*. Switching state *S1 (S3)* and *S2 (S4)* with intermediate switches SW5 (SW7) and SW6 (SW8) respectively defines a 4 × 4 switch. This 4 × 4 switch shown by bounded box in Figure 1 forms the *X-Structure*.

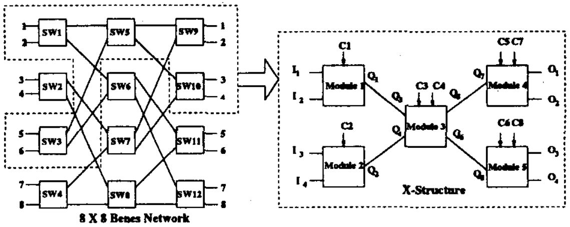

Figure 1. An 8×8 **Switch and X-Structure**

Table 1. Connectivity Diagram of 8×8 **switch**

Switching State	Input Ports	Output Ports	Intermediate Switch
S1	1,2,5,6	1,2,3,4	SW5
S2	1,2,5,6	5,6,7,8	SW6
S3	3,4,7,8	1,2,3,4	SW7
S4	3,4,7,8	5,6,7,8	SW8

The *X-Structure* provides the basic switching functionality which is exploited along the temporal and spatial dimensions to obtain higher order switch interconnects. The following section describes in details the *X-Structure* and its use by interleaving the *X-Structure* in space, and multiplexing in time to achieve complete switch functionality of a 8 × 8 switch.

3. The X-Structure Based ATM Switch

The *X-Structure* (See Figure 1) forms the basic building block of the switch. Modules 1 and 2 are $2-1$ multiplexers, modules 4 and 5 are $1-2$ demultiplexers with multiple fan outs and module 3 is a 2×2 switch with multiple fan outs. The multiple fan outs help achieve multicasting. Module M3 performs the function of intermediate switches SW5, SW6, SW7, SW8.

Let τ_c denote the critical delay in switching packets from input ports $I1 \ldots I4$ (Figure 1) to output ports $O1..O4$ and T denote the period of the system clock. A hardware reuse factor is defined as $\rho = \left\lfloor \frac{T}{\tau_c} \right\rfloor$ such that ρ is the nearest multiple of 2. Thus in general, the *X-Structure* can be temporally pipelined to an extent bounded by ρ. The *Super-channels* can thus be reused to an extent bounded by ρ. For a 0.25μ technology, $\rho = 2$.

Table 2. Details of X-Structure Controls

Module 1&2			Module 3			Module 4			Module 5		
C1	C2	Connec -tivity	C3	C4	Connec -tivity	C5	C7	Connec -tivity	C8	C6	Connec -tivity
0	0	$I_1 \rightarrow Q_1$ $I_3 \rightarrow Q_2$	0	0	$Q_3 \rightarrow Q_5$ $Q_4 \rightarrow Q_6$	0	0	$Q_7 \rightarrow O_1$	0	0	$Q_8 \rightarrow O_3$
0	1	$I_1 \rightarrow Q_1$ $I_4 \rightarrow Q_2$	0	1	$Q_3 \rightarrow Q_6$ $Q_4 \rightarrow Q_5$	0	1	$Q_7 \rightarrow O_2$	0	1	$Q_8 \rightarrow O_4$
1	0	$I_2 \rightarrow Q_1$ $I_3 \rightarrow Q_2$	1	0	$Q_3 \rightarrow Q_5$ $Q_3 \rightarrow Q_6$	1	0	$Q_7 \rightarrow O_1$ $Q_7 \rightarrow O_2$	1	0	$Q_8 \rightarrow O_3$ $Q_8 \rightarrow O_4$
1	1	$I_2 \rightarrow Q_1$ $I_4 \rightarrow Q_2$	1	1	$Q_4 \rightarrow Q_5$ $Q_4 \rightarrow Q_6$	1	1	Power Down	1	1	Power Down

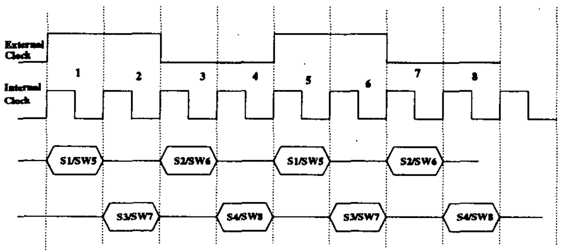

Figure 2. Two Phase Operation of X-Structure

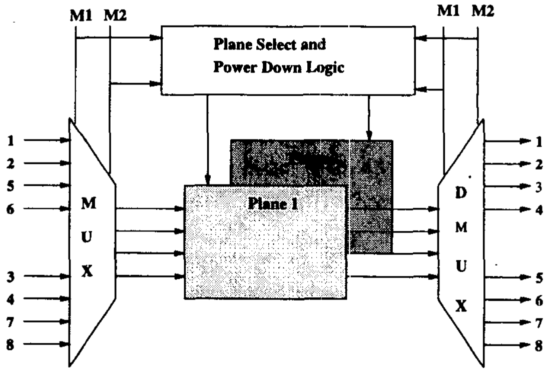

Figure 3. 8 × 8 **Switch using X-Structure**

By using level sensitive logic and two phase clocking [14], the X-structure is temporally pipelined to support the functionality of the two intermediate switches. $(C_1 \ldots C_8)$ form the control set for the various modules. The association of control signals with the various modules and their corresponding functions are given in Table 2. The timing diagram with two phase clocking to achieve 4 × 4 switching using *X-Structure* is shown in Figure 2. The labels S_n/SW_k in Figure 2 corresponds to switching state S_n using intermediate switch SW_k. Appropriate $< C1 \ldots C8 >$ controls are set to achieve this switching function. Thus in one external clock cycle, a complete 4×4 switching operation is established.

3.1. Configuring Higher Order Switches

A 8 × 8 banyan switch is obtained by spatially pipelining two planes of X-Structures and using multiplexer and demultiplexers pairs as shown in Figure 3 with appropriate multiplexer control as in Table 3. The multiplexers in this case serves the dual purpose of multiplexing appropriate set of input ports on to the appropriate switching plane. Note that the two planes (shown in Figure 4) lag each other by $\frac{\pi}{2}$ to avoid bus contention at the output thereby achieving spatial pipelining. Note that level sensitive multiplexers have the property that they are transparent for the duration the control signal is high and latch on the data until the control level is pulled low [14]. Because of these level sensitive multiplexers, we can avoid using latches at the output of the switch. Also a careful design of the switch taking into account the multiplexer delays (approximate 0.2ns), data can be made available at the output in synchronization with the external clock thereby avoiding the use of latches. The power down logic is used to power down appropriate *X-Structure* which are not in use. This is done using an activity sensor. The activity sensor monitors the traffic on the input ports and on detecting an idle plane sets M1 and M2 to 1. The activity sensor then instructs the power down logic to shut down the appropriate plane by assigning appropriate values to control signals $< C1 \ldots C8 >$ (as shown in Table 2. The *X-Structure* can be powered down to standby model by setting the *Power Down* bits (shown in Table 2) to 1. The power down mode is a

Table 3. Details of Multiplexer Control Signals

Multiplexer						Demultiplexer		
M1	M2	Clock Cycle	Clock Phase	Plane Selected	Switching State	Clock Cycle	Clock Phase	Plane Selected
0	0	1	I	P1	S1	1	I	P1
0	0	1	II	P1	S1	1	I	P1
0	0	2	I	P1	S2	2	II	P1
0	0	2	II	P1	S2	2	II	P1
0	1	3	I	P2	S3	3	I	P1
0	1	3	II	P2	S3	3	I	P2
0	1	4	I	P2	S4	4	II	P2
0	1	4	II	P2	S4	4	II	P2
1	1				Reserved			Reserved

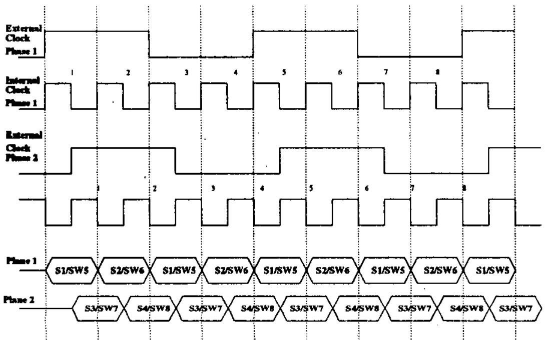

Figure 4. Timing diagram of 8×8 switch

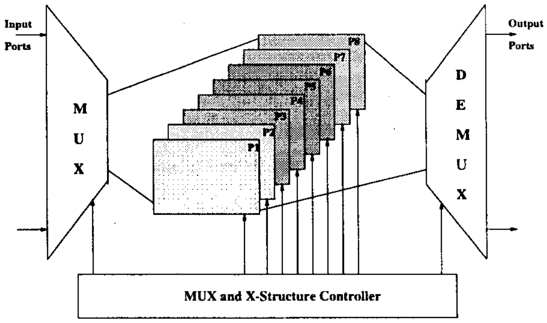

Figure 5. Scalable ATM Switch (32×32)

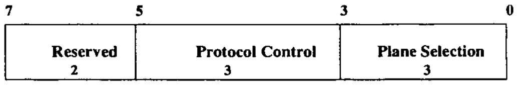

Figure 6. Control word format for multiplexer controls

unique feature that helps achieve low power. This paper does not go into the details of these design issues and we restrict the discussion at an architectural level of the switch design.

Scalability of the switch is limited by the timing constraint of the multiplexer-demultiplexers pair (contention at the output). For a 0.25μ technology, a system clock of T and multiplexer-demultiplexers delay of t_{mux}, the scalability of the switch is limited to atmost $\lfloor \frac{T}{t_{mux}} \rfloor$ planes of X-structures. For the purpose of illustration Figure 5 shows a 32×32 switch supporting eight *Super Channels*. The switch is controlled using a 8 bit *Multiplexer Control Word* shown in Figure 6. The three *LSB* bits are used for selecting the appropriate X-Structure, and the remaining bits are specific to ATM protocol and are not discussed in this paper. Details of this can be found in [12].

3.2. Switch Application

The proposed switch architecture is used at the core of the ATM communication controller. The packet processor used in the ATM communication controller is given in Figure 7. The output of the packet processor (marked as output in Figure 7) forms the input to the proposed *X-Structure* based switch. The communication controller architecture is targeted for realtime VBR multimedia applications. Two separate queues, one for VBR/CBR and the other for ABR/UBR traffic are maintained so that priorities can be assigned to the packets. The queues are implemented as circular buffers. Depending on the type of incoming packet, it is enqueued in one of the queues. The call admission controller used in the communication controller is based on the *Entropy* model proposed in [11]. The *Entropy* model is implemented using a *leaky bucket* mech-

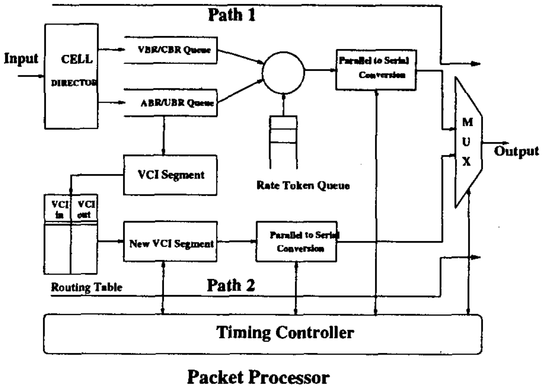

Figure 7. ATM Communication Controller and Packet Processor

anism realized as a counter. The VCI field in every incoming packet is used to hash into a VCI table and the corresponding values of the VCI are changed to that of the outgoing VCI. The ATM cell header which comprises VCI fields and data are processed in parallel, along path 2 and path 1 respectively to reduce the switch latency. Any application that wishes to establish a connection specifies the source and destination address. The timing controller covers these source and destination address to appropriate *X-Structure* and multiplexer control signals as given in Table 3 and Table 2 respectively. This communication controller is used at the *Harmony Layer* of the HARMONY QoS architecture proposed in [10]. The communication controller comprises number of counters, memory management unit, buffers and latches. The communication controller is implemented in VHDL and simulations are carried out to verify its functionality.

4. Performance of the Switch

The design space of the switch is explored with respect to algorithms, architecture, and switch element. We summarize the key results :

- *Performance of the switch vis a vis Algorithms :* A traffic analysis carried out on a 26 node network [13] indicates that on an average the ports remain idle for 28% of the time. Out of this 17% of the time mutually exclusive set of input ports remain idle. Thus during these idle periods, the appropriate plane can be powered down to standby mode to achieve low power.

- *Throughput :* At a traffic load of 0.8, throughput of the switch is 0.86, indicating a maximum utilization of the switching hardware. The aggregate throughput of a 8 × 8 switch is 2.66 Gbps. Details of variation in throughput with traffic load is reported in [15].

- *Latency :* The X-Structure switching element implemented as a n-mos pass transistor logic, was simulated for delay estimates using *Path Mill*[1]. Simulations indicated a switching latency of *i.e* 3.0ns. Details of the variation in latency with traffic load are reported in [15].

- *Slot Stealing :* For a bursty traffic, with traffic loads varying from 0.7 to 0.9 the probability of idle slots across input ports is 0.186. These slots can be used by other active input ports out of turn, thus ideally reusing the hardware 18.6% of the time. However in our implementation, slot stealing can occur at only those input ports having mutually exclusive path sets. Thus the amount of hardware reuse will be less than 18.6%.

- *Performance vis a vis Architecture :* The basic architecture of the X-structure is obtained by partitioning a conventional 8 × 8 switch into independent mutually exclusive path sets. In conventional switch design, due to the absence of temporal and spatial pipelining, only one stage of the switching network is active. Thus for a n stage switching network, the overall hardware utilization is $\frac{1}{n}$. The remaining (n-1) stages of the switch are in standby mode thus consuming standby power of $\frac{n-1}{n} \times P$, where P is the total standby power of the n stage switch. As temporal and spatial pipelining is implemented in X-structure, all the stages of the X-structure are utilized, thus leading to a 100% hardware utilization, and a standby power saving of approximately $\frac{n-1}{n} \times P$. In our implementation n = 3, giving a standby power saving of 66.66%.

- *Power Saving :* Consider a *three* stage 8 × 8 switching element implemented as a *Clos Network* comprising 12 switching elements. The switching elements of the conventional and X-Structure based switches are implemented in n-mos pass transistor logic with a supply voltage of 3.3 volts. in 0.25μ technology. Simulations were carried out using *Power Mill* [2] to estimate the current sourced. Simulations indicated 172 μA current drawn for a conventional switch element and 221 μA for that of a X-Structure switching element. Let P1 be the power consumption of the conventional switch and let P2 be the power consumption of an equivalent switch using X-structures with $\rho = 2$. Then $P1 = 172 \times 10^{-6} \times R \times 12$ and $P2 = 221 \times 10^{-6} \times R \times 4$ where

[1] Path Mill is a SynopsysTM tool

[2] Power Mill is a Synopsys^{TM} tool

Table 4. Area Estimates of ATM Communication Controller

Name	Area in μm^2
Queues	1043780
Counters	327808
VCI/VPI Tables	133907
Cell Processor	1689052.50

R is the load associated with the circuit resulting in a power saving of approximately 2.334 times that of conventional switching fabric. In reality the power saving is less than 2.334, since additional power is required to drive the multiplexer-demultiplexer pair.

- *Area Estimates :* The *X-Structure* and the communication controller are implemented in VHDL using 0.25μ *Texas Instruments* cell library. The area estimate of the *X-Structure* is 1650.40 μm^2. The area estimates of the communication controller is given in Table 4.

5. Conclusion

In this paper we have presented a methodology for reconfigurability and power optimization of ATM switches at all levels of switch design. The *X-Structure* based ATM switch provides dynamic re-routing of cells from active ports through idle ports. Multiple *X-Structures* are spatially pipelined and temporally multiplexed to obtain higher order switches. The limits on scalability of higher order is technology dependent and is determined by the delay of multiplexer-demultiplexer pair.

From a topological analysis of 8 × 8 switching element it is clear that the proposed switch architecture has a 100% hardware utilization as opposed to conventional switch architectures that achieve a hardware utilization of 33.33%. A 8 × 8 multistage switching network was constructed using this switching element. The total silicon area for the communication controller is 3.25 square millimeters in 0.25μ technology. Analysis of this multistage switching network indicates a power saving of 66.66% in standby mode as compared to conventional switch design. Based on traffic analysis it is observed that any input port of the switch is idle 17% of the time. By incorporating power down features an additional power saving of 17% can be achieved. A mechanism of slot stealing is provided to utilize the idle slots. The idle slots are reused 18.6% of the time resulting in enhanced throughput of the switch. The 8 × 8 switch supports a maximum aggregate throughput of 2.66 Gbps.

References

[1] "ATM Forum Standards Document", 1997 *(http://www.atmforum.org/specs/accept/).*

[2] F.A. Tobagi, "Fast Packet Switch Architectures for Broadband Integrated Services Digital Network", *IEEE Proceeding* Vol. 78, 1993, pp. 133-167

[3] Ra'ed Awdeh and H.T. Mouftah, "Survey of ATM switch Architectures" *Computer Networks and ISDN Systems*, Vol.18, November-1995, pp. 1567-1613.

[4] J. Turner and N. Yamanaka "Architectural Choices in Large Scale ATM Switches", *IECIE Transactions on Communications*, Vol. E81-B, January-1998, pp.120-137.

[5] E. Bastruk *et al*, "Design and Implementation of QoS Capable Switch Router", *IBM Research Report*, Technical Report Number : RC 20248 (31-05-99).

[6] M. Katavenes, D. Serpdhos and P. Vlsolaki, "ATLAS-A General Purpose, Single Chip ATM Switch with Credit-Based Flow Control", *Proceedings of the 4th Symposium on Hot Interconnects*, Stanford, California, USA, August 1996.

[7] H.J. Chao, B.S. Choe, J.S. Port, and N. Uzui, " Design and Implementation of Abacus Switch - A Scalable Multicast ATM Switch", *IEEE Journal of Selected Areas in Communication*, June-1997, Vol. 15, pp. 830-843.

[8] Y.Y.Yeh, M.C.Hluchyj and A.S. Acompora, "A Knockout Switch: A simple modular architecture for high performance packet switching" *IEEE Journal of Selected Areas in Communications*, Vol.5, August-1987, pp. 1274-1283.

[9] B.Patel, F. Schaffa and W. LeMair, "The Helix Switch: a single chip switch design", *Computer Networks and ISDN Systems*, October-1996, Vol.28, pp-1791-1807.

[10] Abhijit M. Lele and S.K. Nandy, "Harmony - A Framework for Providing Quality of Service in Wireless Mobile Computing Environment" *Lecture Notes on Computer Science : LNCS Nos 1745*, pp. 299-308.

[11] Abhijit Lele and S.K. Nandy, "Can QoS Guarantees be Supported for Live Video Over ATM Networks", *Proceedings of the IEEE Conference on Global Communications*, Sydney, Australia, November-1998.

[12] Abhijit M. Lele, S.K. Nandy and D.H.J. Epema, "Harmony - An Architecture for Providing Quality of Service in Mobile Computing Environment", *To Appear - Journal of Interconnection Networks*, 2000.

[13] Sullivan E.D, "P-NNI Draft Specifications", *ATM Forum* 94-0471R3.

[14] A.T. Ishi, "Timing in Level-Clocked Circuits", PhD thesis, Massachusetts Institute of Technology, 1991, *(ftp://lcs.mit.edu/MIT/LCS/TR-552).*

[15] Abhijit M. Lele and S.K. Nandy, "Performance Evaluation of Low Power Gigabit ATM Switch", *Proceedings of the Asia Pacific Conference in Communications*, Seoul, Korea, October-2000.