Survey of Switch Architectures

Purvesh B. Thakker

ECE 412 - Computer Architecture
University of Illinois at Urbana-Champaign
Professor Constantine Polychronopoulo
December 11, 1998

Abstract

This year more Internet traffic will traverse worldwide networks than voice traffic, and Internet traffic will continue to double every 100 days. With such dramatic traffic increases, the network infrastructure will naturally need to evolve to handle orders of magnitude more traffic. With optical fiber technology evolving rapidly, network switches have become the bottleneck in the network infrastructure. This paper evaluates the requirements of an Internet Age switch and then surveys a few of the different approaches to switch design including the Crossbar, Knockout, Shared Memory, and Sunshine Switches. This paper finally concludes that all architectures will perform well for small-scale switches, but Banyan-based switches possess the most potential for large-scale telecom needs.

Table of Contents

1. Introduction

This year more data traffic will traverse worldwide networks than voice traffic [Ullal98]. Some people say that Internet traffic is growing at 70% per year, and others say that it doubles every 100 days. Either way, the Internet is growing at a rapid pace. This growth comes both from the increase in popularity of the Internet and the fact that the absolute number of computers is growing at 20% - 30% per year [Sindhu98]. It is obvious that the underlying network infrastructure needs to grow with the traffic traversing it. What parts of the infrastructure need to be upgraded? Networks currently exist that will send over 2 Gbits of information per second. In research labs, fiber optic cables have carried 80 Gbits per second. Experts say that this number will increase to 300 Gbps by the year 2000 [Sindhu98]. The incredible statistic is that this is still only 1% of the capacity of a single fiber! These advances illustrate that raw point-to-point bandwidth is not the bottleneck that will arise.

Moving up the line we find electronic devices such as switches, routers, and bridges. These devices inherently cannot keep up with their optical fiber counterparts because photons simply move through glass faster than electrons move through metal. These electronic devices must evolve to handle orders of magnitude more traffic. This paper surveys current research being done in the area of switch architecture. First, switch requirements will be outlined. Next, the basic switch architecture and more specific architectures will be presented. Finally, some more revolutionary ideas are presented.

2. Switch Requirements

Telecom companies have traditionally used a circuit switching approach. This approach consists of a long-term fixed bandwidth connection and is useful to transport stream-type traffic such as voice and video. The approach, however, is not well suited to the more erratic traffic patterns of data. Packet switching evolved as a more efficient way to carry data communication traffic. This approach allows dynamic sharing of communications resources and typically involves buffering, variable throughput, and variable delay [Ahmadi89].

With the packet switching paradigm, a switch is a multi-input, multi-output device, and its job is to get as many packets as possible from the inputs to the appropriate outputs. A switch of the Internet age should address the following needs:

- Maximized throughput
- Scalability
- Stackability
- QoS capability
- Multicasting
- Congestion control

The need for high throughput was illustrated in the introduction. Switch throughput is defined as the maximum sustained throughput that the switch can handle summed across all inputs, and the more traffic that a switch can handle, the better. Scalability is another major concern. Large-scale ATM switches are widely recognized as an important component in building the Information Superhighway. ATM switches

with hundreds of input and output ports and gigabit link rates have been built in laboratories [Lin98]. The reason that scalability is so important is that telephone switches need to service millions of customers. Telephone switches today commonly have tens of thousands of ports. Stackability is a valuable feature that allows smaller customers to add bandwidth as the need arises. With additional ports, the bandwidth of the switch has to be increased. In other words, a stackable switch does not simply consist of linking two switches together. A single switch should ideally have full bandwidth from each input to separate outputs simultaneously. Quality of Service (QoS) capabilities are widely recognized as an important part of future networks. ATM QoS priority can be set by two parameters: 1) timing or delay, and 2) protection or cell loss rate. Multicasting is another valuable feature. Rather than having five copies of video come through an input and then split up to five output ports, one copy will go to the switch and the switch will send copies to each desired output port. QoS and multicasting are relatively new ideas that haven't yet made it into industry but are widely recognized as important features of an Internet Age switch.

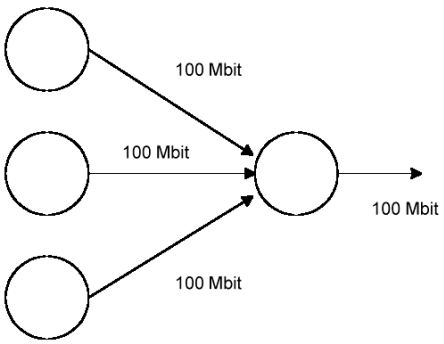

Figure 1: Output Congestion

Congestion control is an important task that a switch must deal with. Congestion occurs under several different scenarios:

- When traffic at different inputs merges to the same output, all of the traffic will not be able to get through right away.
- When a rate mismatch occurs
- When there is a rate mismatch in broadcast

Intentional loss of packets is not a desirable alternative to congestion control. If a packet is dropped, then the receiver will need to request a retransmission and will then ignore any other packets received until the desired packet arrives or a timeout occurs. If these retransmissions are dropped, then the receiver will not request another retransmission until it times out. With this scenario, throughput nearly comes to a halt. The switch should be designed to minimize dropped packets in order to avoid multiple retransmissions. Switches are designed to maximize network efficiency, not switch efficiency. Any packet loss is detrimental to this goal.

On a final note, evaluating the performance of switches turns out to be a serious challenge. The throughput of a switch is a function of the traffic to which it is subjected. There are several elements to a traffic model [23]:

- When do packets arrive
- What outputs are they destined for
- How big are they

A good traffic model will not assume that traffic is random since that does not reflect reality. Traffic will often flow in bursts from a given input or to a given output. Traffic between different inputs and between different outputs will also be correlated.

3. Basic Switch Layout

A switch consists of input ports, output ports, and a fabric. First the input ports receive packets from the outside world. Next, the fabric delivers the packets to the right output port. Finally the output ports deliver the packet back to the outside world. The ports deal with the complexity of the real world so that the fabric can do its simple job. Before a packet is passed from the input port to the fabric, the input generally figures out where the packet needs to go. It then either sets up the fabric to do its job, or passes enough information to the fabric so that the fabric can self-route the packet.

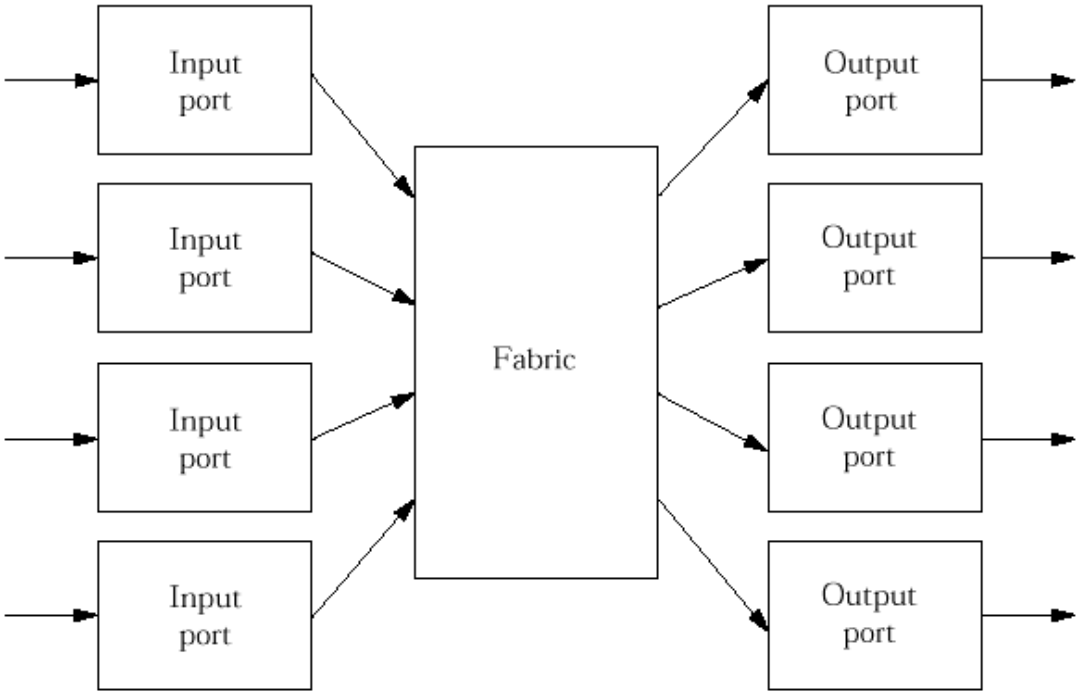

Figure 2: Basic Layout of a Switch [Peterson 191]

Another key tool of a switch is buffering. Buffering can happen in the input port, the output port, or even in the fabric. Buffering helps alleviate congestion problems, but cannot eliminate them completely. If several inputs attempt to send a stream of traffic to a single output, then eventually the buffers will fill up and packets will have to be dropped. Input buffering has a serious limitation known as Head-of-Line (HOL) blocking. If output congestion occurs, then some packets will be stuck in input buffers preventing the packets behind them from moving to other outputs. When traffic is uniformly distributed among outputs, head-of-line blocking limits throughput of an input-buffered switch to 59% of the theoretical maximum (sum of link bandwidths) [Karol87]. Input buffering is still a useful tool, however, since techniques are available for

alleviating HOL blocking. For example, the input buffers can be split up according to output in order to create several HOL queues [Kolias98, Son97]. Using output buffering or shared buffering permits higher throughputs than input buffering schemes, but even output buffering has it challenges. Often output buffering requires a complex architecture [Hui90] or internal operation at several times the system speed [DePryck91]. Most real switches will use a combination of input and output buffering.

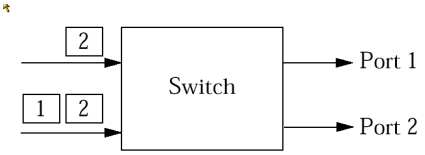

Figure 3: HOL Blocking [Peterson 192]

4. Specific Switch Architectures

Thousands of papers have been published in the area of switch architectures presenting a wide variety of approaches and a wide variety of variations on these approaches. The following Switches provide excellent insight into many of these architectures:

- Crossbar
- Knockout
- Shared Memory
- Batcher-Banyan Fabrics
- Sunshine Switch

4.1 Crossbar Switch

Crossbar switches are widely used in traditional telecom voice applications. In a crossbar switch the only type of contention is output contention since every input is connected to every output. The output port of the fabric plays a critical role in routing the packet. It has two functions:

- Recognize the packets that are destined for its outputs
- Deal with contention when multiple packets go to the same output.

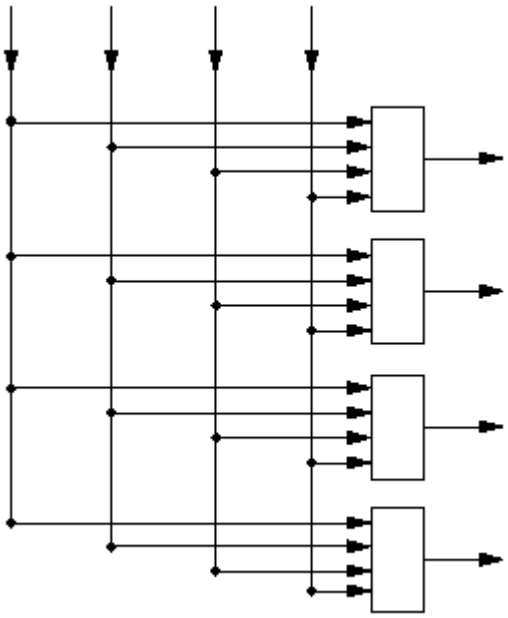

Figure 4: Crossbar Switch [Peterson 192]

Scalability is a major drawback of crossbar switches. With an increase in the number of ports, the size of the crossbar fabric must increase almost exponentially. The complexity of the output port grows in proportion to or faster than the number of inputs

n. Since there are also n outputs, the complexity of the entire switch grows at least as fast as n^2. One modification that can be made to the Crossbar Switch involves turning it into a hybrid crossbar. Just like using a switch port to link another switch, the bottleneck becomes the interconnect link. The Crossbar switch also requires that all ports operate at the same speed. To match dissimilar speeds such as 10/100/1000 Mbps Ethernet, switch designs would get complicated quickly. A typical switch today might have 24 ports running at 100 Mbps and 2 Gigabit ports that link to other switches and routers. The crossbar has very good point-to-point switching characteristics such as in traditional telecom applications, but it is not well suited to more flexible computer networks. With bursty computer traffic, too many packets would be lost due to output congestion.

4.2 Knockout Switch

The Knockout switch architecture is attractive for large-scale switch implementations because of good cell loss performance. The Knockout Switch is an enhanced version of the Crossbar Switch. Some traffic assumptions are made to reduce the complexity of the output ports. The Knockout Switch has an output port that can accept *l* packets simultaneously with *l* less than the number of ports. The output port of the Knockout switch has three parts:

- A set of packet filters that recognize packets destined for that port.
- The "knockout" system, called a concentrator, which selects up to *l* packets from those destined for the port, and discards any excess packets. (hopefully rarely)
- A queue that buffers *l* packets

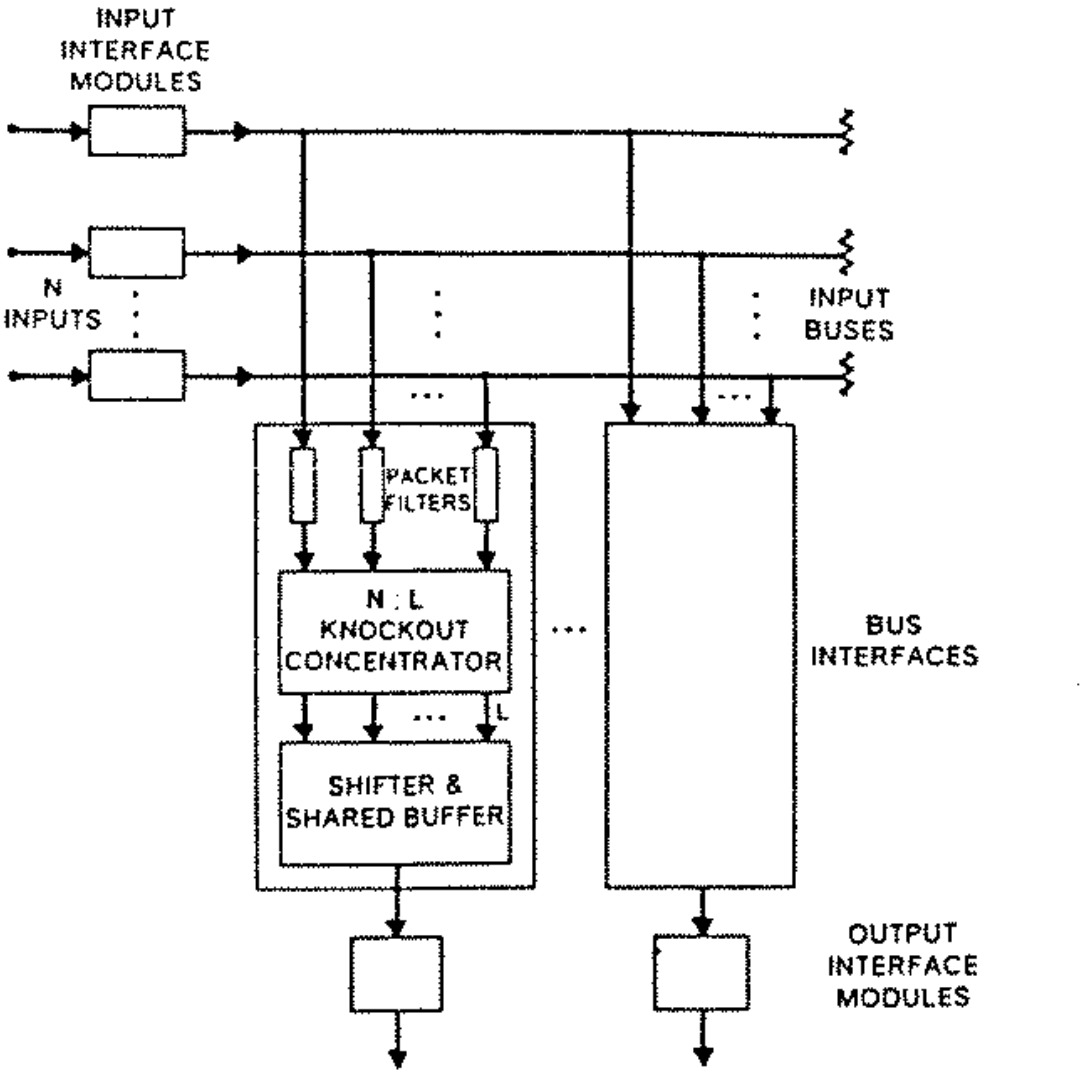

Figure 5: Knockout Switch [Ahmadi89]

The filters use simple matching hardware to identify packets that contain the right output port number. The concentrator must be fair. Fairness is achieved by playing the packets against each other in a knockout "tournament." The knockout tournament selects *l* winners from *n* contestants. The first section is the traditional knockout tournament,

which selects one winner. All the losers from that section move on to compete to be runner up with this consolation tournament running in parallel with the first tournament. This continues until *l* packets have been selected. The output ports have a special complication to deal with due to the output buffering. In one cycle, they must be able to accept up to *l* packets and transmit one packet. Instead of a FIFO that runs *l* times faster than the output, an array of *l* buffers is used preceded by a shifter. This way the output port may accept a packet into each of the *l* buffers every cycle, and preserve ordering by drawing from each buffer in turn.

With respect to scalability, the knockout switch grows more slowly than the Crossbar. The parameter *l* stays fixed at some value, so the buffers do not need to increase. The complexity of the concentrator for a large *n* value approaches *n* x *l*, so it scales with *n*. The number of packet filters per port is *n*. Thus the complexity of the output port (mostly concentrator) is proportional to *n*. Since there are *n* output ports, the total switch complexity is better than n^2. The Knockout switch deals with output congestion better than the Crossbar switch also. It can accept up to *l* packets in one cycle. Early on switch designers assumed that traffic going to outputs was uncorrelated, but this turns out not to be true. The value of *l* should be chosen to account for simultaneous bursts of traffic from different inputs. One variation of the knockout switch is the Distributed-knockout Switch. An input port with packets waiting sends a cell to the switch at the beginning of each time slot. The cell sent out from an input port reaches either its destination output port (wins contention) or a different input port (loses contention). The advantage of the distributed knockout switch is even better output congestion control. The switch only drops cells which lose contentions a set consecutive

number of times. With a simple priority scheme, this switch is also capable of preserving cell sequencing [Cheng96].

4.3 Shared Memory Switch

Shared media switches work by using a shared medium in place of the fabric. The nice thing about shared memory switches is that the whole switch is one big buffer that can be built out of off the shelf memory chips. It can be made to work like a perfect crossbar, but better, since the buffering is shared among all of the output ports. This class of switches tends not to scale very well, though, because the shared resource either gets more overloaded or needs to get faster as the switch size grows. The memory bandwidth required grows rapidly as the number of ports increases. At a high port count, the bandwidth becomes unreasonably high because the bus connecting the memory must run *n* times faster than the line speed of a single link. Still, it is possible to widen the memory path. It turns out that the main limitation on size for a switch like this is the rate at which the control logic can operate. Shared memory approaches depend on a central switch engine to provide high-speed interconnection to all ports. Unlike the memory bandwidth, the control logic cannot be sped up through widening.

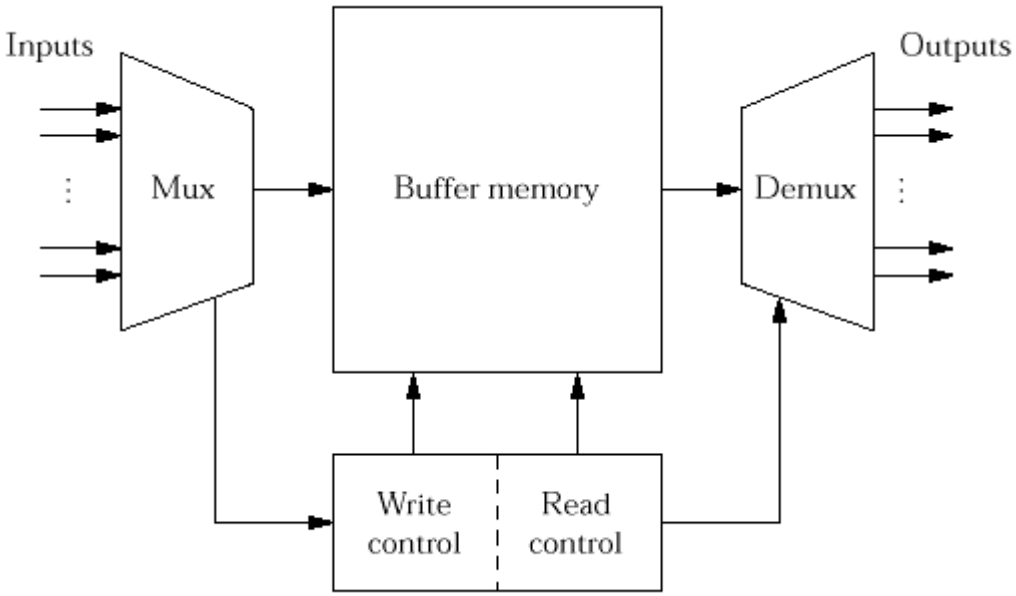

Figure 6: Shared Memory Switch [Peterson 201]

4.4 Batcher-Banyan Fabric

Self-routing fabrics are the class of switches that have achieved scalability best. Self-routing fabrics are made up of many small interconnected switching elements. The packets find their own way through the fabric based on a sequence of local switching decisions made at each small switching element. The principle behind self-routing fabrics is that each packet carries enough information in its header to allow the small switching elements to make a local decision. The input usually adds an extra internal header to the packet before sending it on to the fabric, and the output port strips this internal header before sending it out. Self-routing fabrics are often made from simple 2 x 2 switching elements that switch based on just one bit in the self-routing header.

The Banyan fabric is one such self-routing fabric. Each element looks at the most significant bit of the internal header and routes packets to the top or bottom output based

on whether the bit is “0” or “1.” Looking at the entire Banyan fabric in figure 7, it can be seen that the first column routes packets to the correct top or bottom half of the network. Then the next column routes the packet to the correct quarter of the network. The clever part is the way that the switches are arranged to avoid collisions. The Banyan network will avoid collisions if two restrictions are placed on the inputs. First, the inputs should be arranged in ascending order. Second, there should not be more than one input heading to the same output. These are severe restrictions that can be relieved with techniques that will be discussed later. Compared to the Crossbar, Knockout, and Shared Media switches, the Banyan network has a pipelining effect that improves throughput. Once the first set of inputs passes through to the second column, the next set of inputs may enter the first column. Thus a packet may move closer to its output without blocking the path of any other packet. With respect to scalability, a Banyan network with n inputs needs to have $\log_2 n$ stages. Each stage has $n/2$ switching elements. Thus the complexity of the switch grows as $n\log_2 n$, a good growth rate. The Banyan network is also the simplest switch matrix architecture that can be used as a building block for more complex switch matrix architectures such as the Bene matrix. [Okayama94].

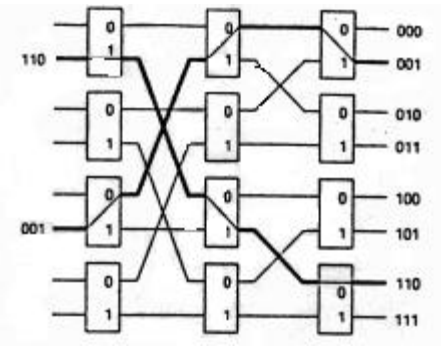

Figure 7: Banyan Network [Ahamdi89]

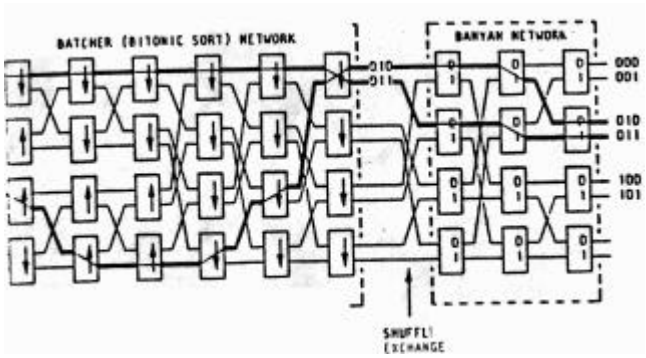

Figure 8: Batcher-Banyan Fabric [Ahamdi89]

The Batcher matrix addresses the first restriction of the Banyan network, ascending inputs. The fabric sorts the packets before sending them on to the Banyan network.. There are two types of elements to the Batcher fabric, those that sort up, and those that sort down. Figure 8 shows how the packets are sorted. The Batcher network essentially implements a recursive merge sort algorithm in hardware. With Batcher-Banyan networks, the fabric delivers all packets to the right output, as long as there are no duplicate packets heading for the same port. One thing to note about Batcher-Banyan networks is that they use input buffering and so suffer from HOL blocking. Several techniques are available to reduce the blocking as follows:

- Increase the internal speeds relative to the external speeds
- Place buffers in every switch element
- Use a handshaking mechanism between stages or a back-pressure mechanism to delay the transfer of blocked packets.

- Use multiple networks in parallel to provide multiple paths from any input to any output or multiple links for each switch connection [Kruskal83, Kumar86].
- Use a distribution network in front of the Banyan fabric to distribute the load evenly. [Ahmadi89]

4.5 Sunshine Switch

The Sunshine switch adds three elements to the Batcher-Banyan fabric which allow it to overcome its unique output restriction.

- the trap
- the selector
- delay boxes

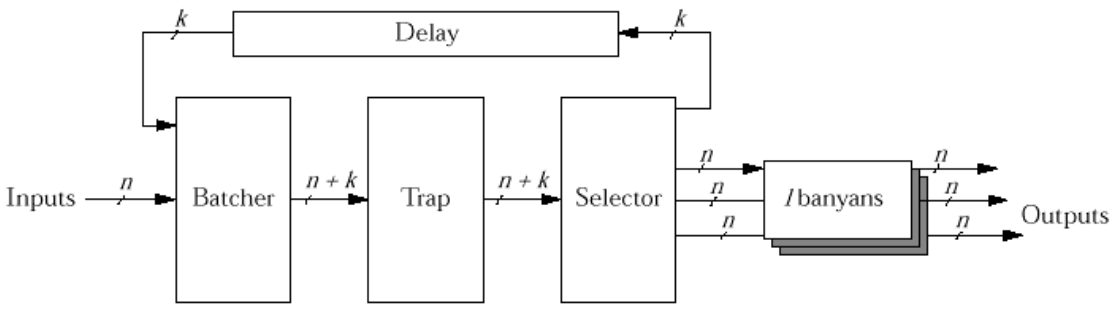

Figure 9: Sunshine Switch [Peterson 200]

The Sunshine network also uses multiple parallel Banyans. When more than *l* packets are destined for the same output in the same cycle, they are recirculated back

through a delay box and resubmitted in the next cycle. The trap identifies up to *l* packets to pass and marks the rest for recirculation. Then the selector selects separate Banyans to use if packets are destined for the same output. The remaining packets are recirculated. Some method of dropping packets must be implemented for when an output averages more traffic than it can handle. Each packet contains a priority field that is incremented each time it is recirculated, and the packet is dropped if this counter gets too high. The Sunshine Switch turns out to be a very good compromise design. It has lower complexity than the knockout switch, and rather than discarding packets, it recirculates them.

5. All-optical Network

A more revolutionary approach towards increasing switch capacity involves creating all-optical networks. In recent years, signal capacity and transmission distance without electrical repeaters have been much improved for optical fiber transmission. Transparent optical switches are attractive for handling such broadband optical transmission lines because they can manipulate them at much higher rates. Effectively, chips would be built using glass to transmit photons instead of metal to transmit electrons. Routing would be done dynamically by changing the frequency of the data received. A switch would have many inputs consisting of multiple fibers that each have many frequencies which are each further time-division multiplexed. Lasers have been demonstrated that transmit 200 Mbps simultaneously at nine different wavelengths [Zirngibl94]. Research shows that up to 400 different wavelengths at 2.5 Gbits per channel could be supported transparently over 1000 km with 50 km amplifier spacing. The implications of this are that a centralized wavelength routing node can provide

interconnection on a national scale. This translates into non-blocked interconnection of at least 20 million customers with an average of 2.5 Mbits per customer [Yu97]. A 634 Gbit switch could be constructed, which is several times larger than current available switching systems using electronics. This system would use a fabric with 6 Tbits of throughput [Robinson96].

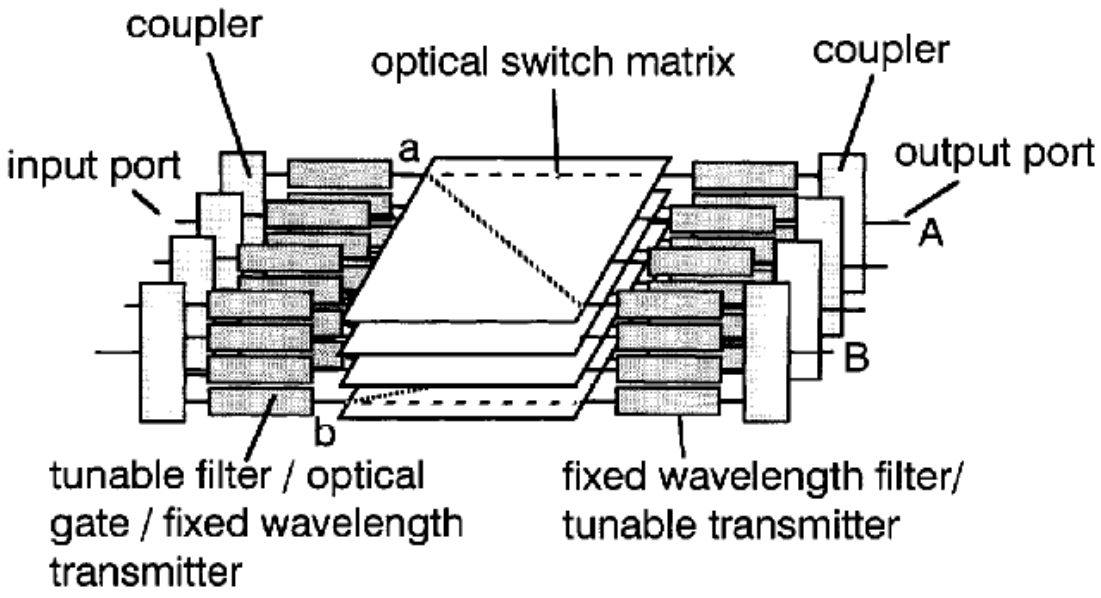

Figure 10: Possible Architecture of All-optical Network [Okayama98]

Frontiernet is an all-optical network that has been proposed. An input module consists of an input buffer, tunable converter, and a combiner. The inputs would buffer the data and reroute the different wavelengths to the frequencies needed to transmit them over the fabric. Two contention resolution schemes are the “store-and-forward” and “hot-potato” schemes. With Store and forward only one of the contending packets is converted to the suitable frequency channel, and the others are stored in the input buffer

and forwarded to the converter in the next time slot. With the hot-potato routing scheme, all circulating packets but one are misrouted to output highways that produce longer paths to the circulating destination. The Frontiernet architecture is promising for use in high capacity photonic switching systems [Sasayama95].

6. Conclusion

We are in the midst of a bandwidth explosion. Combining new network technologies such as ADSL, cable modems, and satellite transmission with the traffic demands of the Internet will provide lucrative incentives to upgrade the network infrastructure. ADSL technology has already begun to catch on bringing 10 Mbps access to the home, a 200 fold increase over today's typical 56kbps modems. The technology involves running fiber to the curb and then using the existing copper wire to get into the home. Cable modems provide similar access speeds using the high-bandwidth television infrastructure. With consumers upgrading to these permanent high performance connections over the next few years, traffic over worldwide networks will reach new orders of magnitude. In 1993 the University of Illinois at Urbana-Champaign generated 40 Mbits of traffic per day. Today, just 5 years later, traffic has jumped to 1 Gbit per second. Universities are, of course, are at the forefront of the world with regards to Internet usage, so the rest of the world will follow similar trends. The networking industry has seen these trends and is gearing up to capitalize on the opportunity. A recent article named four companies including some start-ups that are building the equivalent of 80 pre-Internet AT&T networks. Of course, at the same time the traditional big players are building up their infrastructure. Bill Gates, an Arabic prince, and another billionaire

currently are pursuing a project dubbed Teledesic. The project involves putting hundreds of satellites into space. Homes would connect to these satellites with small dishes, the packets would be routed around in space, and finally the data would be beamed directly down to its final destination. Another creative initiative came out of military technology. The military would often fly planes out over the field and effectively use them as temporary satellites. Somebody recognized that this could be used for Internet access for large cities. Today, at least two companies are following up on the idea. They have designed a special airplane and are targeting 119 metropolitan cities. With all of these initiatives, two scenarios have been floated. One theory is "If you build it they will come." The other theory is that the price of long distance telephone calls will drop to 1 cent per minute in a few years due to fierce competition and companies will start merging to form new alliances. Either way the consumer gets all the bandwidth he can use.

The bandwidth explosion will naturally have a ripple effect on all other related technologies. One of these technologies, switches, will have to evolve to handle orders of magnitude more traffic. Looking at only small switches, just about any architecture will perform about equally well. Larger switches, however, are more sensitive to the architecture chosen [Ahmadi89]. Since telecom companies service millions of customers they will likely pursue the more scalable technologies based on Banyan fabrics. Office and building switches will likely not need to be as scalable, so they will choose the cheapest technology that does the job. Perhaps in the distant future, all-optical networks will provide all of the switching performance that the world could possibly demand. One thing is certain. Researchers and industry are both responding to the challenge of the "World Wide Wait."

7. References

[Ahamdi89] Ahmadi, H., and Denzel, W., "A sof modern high-performance switching techniques", *IEEE Journal on Selected Areas in Communications Vol. 7 no. 7* (September 1989): 1091-1103.

[Cheng96] Cheng, Y.-J., Lee, T.-H., and Shen, W.-Z., “Design and performance evaluation of a distributed knockout switch with input and output buffers,” *IEE Proceedings - Communications Vol. 143, no. 03* (June 18, 1996): 149-154.

[DePryck91] De Prycker, M. Asynchronous transfer mode. Ellis Horwood, Chichester, UK. (1991)

[Karol87] Karol, M.J. Hluchyj, M.G. Morgan, S.P. “Input versus output queuing on a space-division packet switch”, *IEEE Trans. COM-35, 12* 1347-1356 (1987)

[Kolias98] Kolias, C., and Kleinrock, L., “On the odd-even ATM switch,” *Electronics Letters Vol. 34, no. 06* (March 19, 1998): 576-577.

[Kruskal83] Kruskal, C. P., and Snir, M., “The performance of multistage interconnection networks for multiprocessors”, *IEEE Trans. Comput., Vol. C-32* (December 1983): 1091-1098.

[Kumar86] Kumar, M., Jump, J. R., “Performance of unbuffered shuffle-exchange networks”, *IEEE Trans. Comput., Vol. C-35* (June 1986): 573-577.

[Hui90] Hui, J. Switching and traffic theory for integrated broadband networks. Kluwer Academic, Boston,. (1990)

[Lin98] Lin, Y.-S., Shung, C.B., and Chen, J.-C., “Design of knockout concentrators,” *IEE Proceedings - Communications Vol. 145, no. 03* (June 16, 1998): 145-151.

[Okayama98] Okayama, H., and Kamijoh, T., “Blocking property of cross-connect node architecture using star couplers, wavelength converters and optical switch matrices,” *Electronics Letters Vol. 34, no. 10* (May 14, 1998): 1005-1007.

[Okayama94] Okayama, H., and Kawahara, M., “Prototype 32 x 32 optical switch matrix,” *Electronics Letters Vol. 30, no. 14* (July 7, 1994): 1128-1130.

[Peterson] Peterson, L. L., and Davie, B. S., *Computer Networks: A Systems Approach*, San Francisco, CA: Morgan Kaufmann Publishers, Inc., 1996.

[Robinson96] Robinson, A., and O'Mahony, M.J., “Performance of an optical backplane bus for switch interconnection,” *IEE Proceedings - Optoelectronics Vol. 143, no. 04* (August 21, 1996): 237-243.

[Sasayama95] Sasayama, K., “Multihop Frontiernet using generalised perfect shuffle interconnection topology,” *Electronics Letters Vol. 31, no. 13* (June 22, 1995): 1087-1088.

[Sindhu98] Sindhu, P., “Supercharging bits through the Internet”, Siliconindia (October 1998): 42.

[Son97] Son, J.W., Lee, H.T., Oh, Y.Y., et al., “Performance of an input-queued ATM switch with even/odd switching planes,” *Electronics Letters Vol. 33, no. 14* (July 3, 1997): 1192-1193.

[Ullal98] Ullal, Jayshree, “From Dialtone to Webtone”, Siliconindia (October 1998): 34-35.

[Yu97] Yu, A., O'Mahony, M.J., and Hill, A.M., “Transmission limitation of all-optical network based on NxN multi/demultiplexer,” *Electronics Letters Vol. 33, no. 12* (June 5, 1997): 1068-1069.

[Zirngibl94] Zirngibl, M., Joyner, C.H., and Stulz, L.W., “Demonstration of 9 x 200 Mbit/s wavelength division multiplexed transmitter,” *Electronics Letters Vol. 30, no. 18* (September 1, 1994): 1484-1486.