Terabit Burst Switching
Progress Report (12/97-2/98)

Jonathan S. Turner
jst@cs.wustl.edu

WUCS-98-16

June 10, 1998

Department of Computer Science
Campus Box 1045
Washington University
One Brookings Drive
St. Louis, MO 63130-4899

Abstract

This report summarizes progress on Washington University's *Terabit Burst Switching Project*, supported by DARPA and Rome Air Force Laboratory. This project seeks to demonstrate the feasibility of *Burst Switching*, a new data communication service which can more effectively exploit the large bandwidths becoming available in WDM transmission systems, than conventional communication technologies like ATM and IP-based packet switching. Burst switching systems dynamically assign data bursts to channels in optical data links, using routing information carried in parallel control channels. The project will lead to the construction of a demonstration switch with throughput exceeding 200 Gb/s and scalable to over 10 Tb/s.

[0]This work is supported by the Advanced Research Projects Agency and Rome Laboratory (contract F30602-97-1-0273).

Terabit Burst Switching
Progress Report (12/97-2/98)

Jonathan S. Turner
jst@cs.wustl.edu

This report summarizes progress on the Terabit Burst Switching Project at Washington University for the period from December 15, 1997 through March 15, 1998. Efforts during this period have concentrated on working out details of the burst switch architecture, evaluating a variety of implementation alternatives and developing the physical design of the 160 Gb/s ATM switch to allow demonstration of the burst switch within a realistic network context.

1. Fully Scalable Burst Switch Architecture

Given the rapidly growing demand for communications bandwidth (by some estimates, bandwidth usage in the Internet is doubling every six to twelve months), scalability of network elements is essential to keep up with exploding needs. Given the relatively slow rate of improvement in our ability to transmit information serially (due to the well-known opto-electronic bottleneck), the key to scalability in both transmission and switching equipment is the accommodation of large numbers of parallel data channels. In transmission systems, WDM is the key to obtaining the requisite parallelism. Burst switching carries these parallel data channels throughout a switching system and dynamically assigns data to these channels to make the most effective use of the available bandwidth. In addition, burst switching systems will need to support large numbers of ports, in order to achieve the total system capacities that will ultimately be required to keep pace with growing demands.

Figure 1 illustrates a scalable burst switch architecture consisting of a set of *Input/Output Modules* (IOM) that interface to external links and a multistage interconnection network of *Burst Switch Elements* (BSE). The interconnection network uses a Beneš topology, which provides multiple parallel paths between any input and output port. A three stage configuration comprising d port switch elements can support up to d^2 external links (each carrying many WDM channels). The topology can be extended to 5,7 or more stages. In general, a $2k+1$ stage configuration can support up to d^k ports, so for example, a 7 stage network constructed from 8 port BSEs would support 4096 ports. If each port carried 128 WDM channels at 2.4 Gb/s each, the aggregate system capacity could exceed 1,250 Tb/s.

The operation of the scalable burst switch is illustrated in Figure 2. An arriving burst is preceded by a *Burst Header Cell* (BHC) carried on one of the WDM channels dedicated to control information. The BHC carries address information that the switch uses to determine how the burst should be

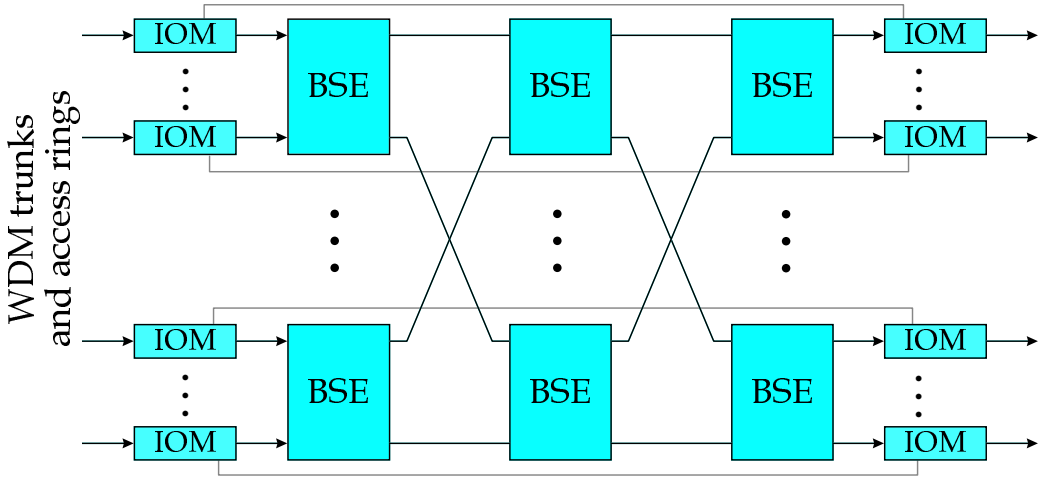

Figure 1: Scalable Burst Switch Architecture

routed through the network, and specifies the WDM channel carrying the arriving data burst. The IOM uses the address information to do a routing table lookup. The result of this lookup includes the number of the output port that the burst is to be forwarded to. This information is inserted into the BHC, which is then forwarded to the first stage BSE. The data channels pass directly through the IOMs but are delayed on the input by a fixed amount of time to allow time for the control operations to be performed.

When a BHC is passed to a BSE, the control section of the BSE uses the output port number in the BHC to determine which of its output links to use when forwarding the burst. If the required output link has an idle channel available, the burst is switched directly through to that output link. If no channel is available, the burst can be stored within a shared *Burst Storage Unit* (BSU) within the BSE. Figure 3 shows how the BSE can be implemented. In this design, the control section consists of an eight port *ATM Switch Element* (ASE), a set of seven *Burst Processors* (BP), and a *Burst Storage Manager* (BSM). The data path consists of a crossbar switch, together with the BSU. The crossbar is capable of switching a signal on any channel within any of its input links to any channel within any of its output links; so in a system with d input and output links and h data channels per link, we require the equivalent of a $dh \times dh$ crossbar. Each BP is responsible for handling bursts addressed to a particular output link. When a BP is unable to switch an arriving burst to a channel within its output link, it requests use of one of the BSU's storage channels from the BSM, which switches the arriving burst to an available storage channel (if there is one).

In the first stage of a three stage network, bursts can be routed to any one of a BSE's output ports. The port selection is done dynamically on a burst-by-burst basis to balance the traffic load throughout the interconnection network. The use of *dynamic routing* yields optimal scaling characteristics, making it possible to build large systems in which the cost per port does not increase rapidly with the number of ports in the system.

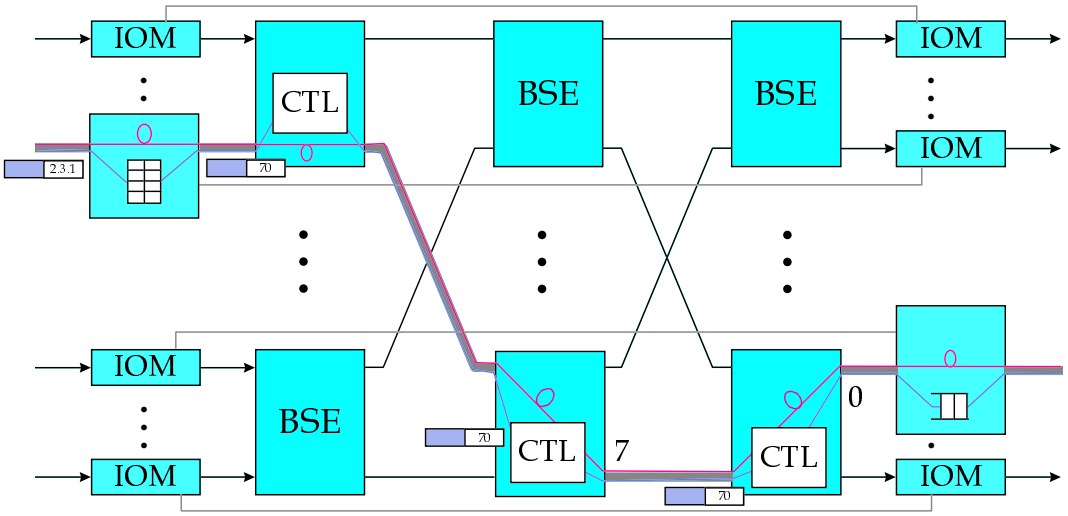

Figure 2: Operation of Burst Switch

2. Timing and Synchronization Issues in Burst Networks

Burst switching systems are intended to accommodate data bursts as small as a kilobyte to over a megabyte. The need to handle short bursts has important implications for how timing and synchronization are handled within burst switching systems.

Figure 4 shows a possible design for a data path through a three stage burst switch. In the diagram, data is retimed at the input IOM and at the output IOM but passes asynchronously through crossbar switches in each of the three stages. As clock frequencies increase, it becomes more difficult to make a design like this work, because at high clock frequencies, the clock period becomes comparable to the uncertainties in the delay between the input and output (uncertainties caused by manufacturing variabilities in the delays of the various components and by the effect of temperature on delays). To make the design work at high clock frequencies, the clocked storage element on the output side can be replaced by a general receiver circuit that extracts timing information from the received data signal to determine the optimum sampling point. However, such receiver circuits require some time to phase lock to the received signal. Consequently, there is a period of time following the establishment of a switched path between input and output IOMs during which data cannot be received reliably.

In burst switches, data bursts may only last a few microseconds and consequently efficient switching of short bursts requires that the receiver circuit be capable of achieving phase lock in no more than about 100 ns. This is a much more demanding requirement than typical receiver circuits can meet. Circuits used in typical commercial data link components operating at gigabit speeds, can require close to a millisecond to achieve phase lock. There is a trade-off between the speed of acquisition and the receiver's stability and bit error rate. Because commercial components typically do not require fast acquisition times, they are optimized to provide stable operation with low bit error

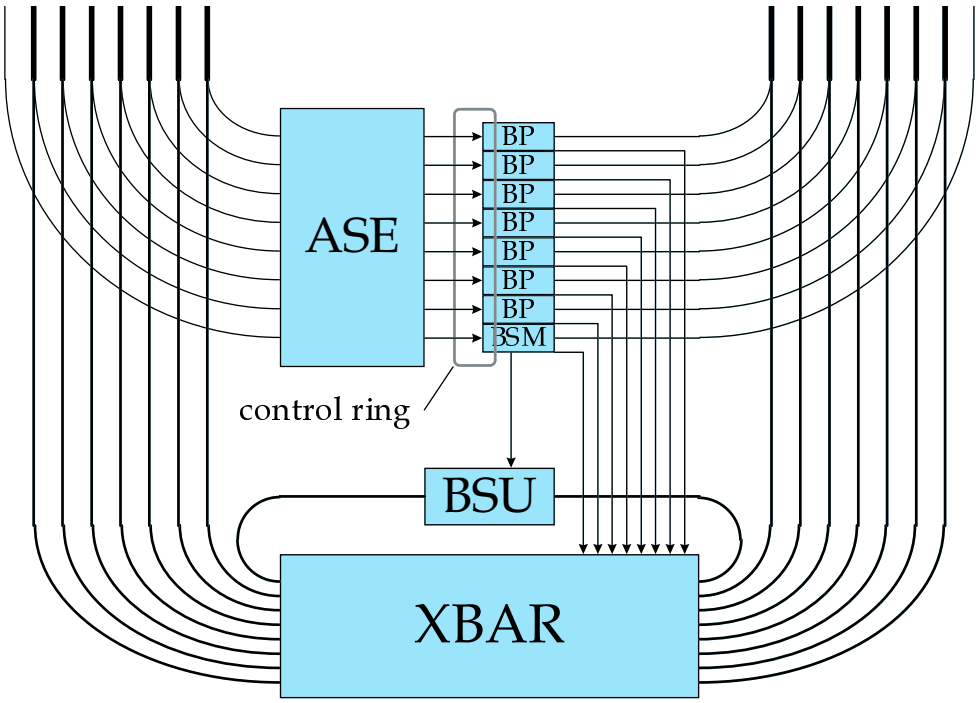

Figure 3: Burst Switch Element

rate. Unfortunately, this makes them unsuitable for use in the burst switching application. Indeed, it is not clear that it is even possible to design circuits that will meet the rapid phase acquisition requirements imposed by burst switching, while achieving acceptable bit error rates.

Fortunately, there is an alternative timing approach that eliminates the need for rapid phase acquisition, which is shown in Figure 5. Here, synchronous crossbars are substituted for the asynchronous crossbars. Each crossbar input contains a phase alignment circuit that adjusts the phase of the incoming data signal to match the local clock timing. Each crossbar output includes a clocked storage element that retimes the data before it is passed to the next stage. Because the timing reference of the data received by the phase alignment circuits never changes, they do not require the ability to rapidly acquire phase lock. Because the phase alignment circuit and the clocked storage element at the crossbar output are all part of the same physical component, the timing uncertainties between these two components are small enough to obviate the need for a complex receiver circuit at the crossbar output, even at gigabit data rates. Still higher data rates can be achieved by reclocking the data within the crossbar or by providing additional parallelism within the crossbar.

It is important to recognize that these issues arise in both electronic and optical implementations of the burst switch data path. It is tempting to think that a burst switch with an optically transparent data path won't have this problem, but that's not really correct. The use of an optically transparent data path just moves the problem from the burst switch to the terminal equipment at the access points of the network, where it becomes even more difficult to handle. With an optically transparent end-to-end path through the burst network, a receiving terminal must be able to rapidly acquire

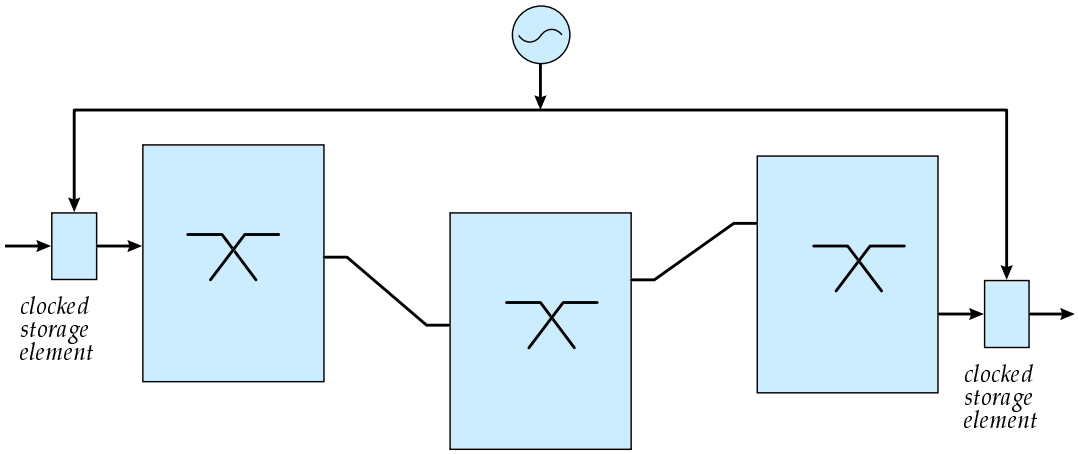

Figure 4: Syncrhonization and Timing with Asynchronous Crossbars

phase information from an optical signal generated by a sending terminal that may be thousands of miles away and connected via a path containing ten or more switches. Just receiving gigabit data with acceptable bit error rates under these conditions is extremely challenging. It's difficult to see how, under these conditions, one could expect to phase lock to the incoming data within the requisite 100 ns needed for efficient handling of short data bursts. The only realistic prospect for operating large burst switched networks using optically transparent end-to-end data paths, involves limiting per channel data rates to a small fraction of what can be achieved with fully synchronous operation and/or greatly increasing the minimum burst size required for efficient operation.

The fact that synchronous operation is essential for high performance burst switching has important implications for the implementation of the burst switch data path. In the short term, the data path can only be implemented electronically, since optical technology is not yet capable of implementing the necessary functions. In the longer term, it is possible that optical technology will mature to the point where the synchronization and timing elements can be implemented in the optical domain and integrated with the crossbar components, so that timing uncertainties can be reduced to the levels required for high speed operation.

3. Technology Assessment for Prototype Burst Switch

We plan to prototype the scalable burst switch architecture described in Section 1. The prototype system will support links with 32 channels per external link (including one control channel) and channel data rates of at least 1 Gb/s. For budgetary reasons, the prototype system will use just a single stage network, but the components will be designed to support the construction of larger scale configurations. Also, for budgetary reasons, our plans do not currently include true WDM links, but rather the use of multi-fiber cables to emulate WDM. Since the principal research issues concern the design of the switching components (and most importantly the control elements), this does not compromise the research objectives in any way.

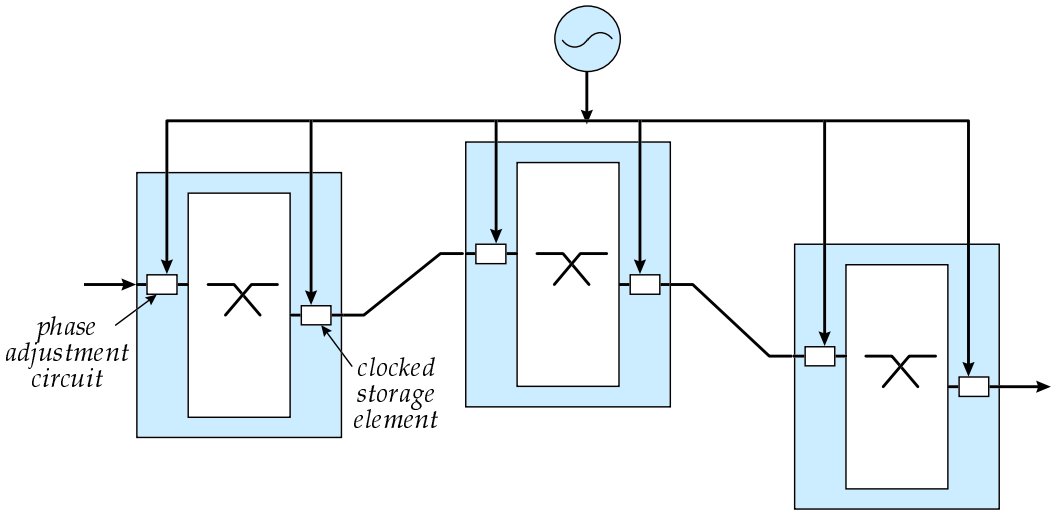

Figure 5: Syncrhonization and Timing with Phase Adjusting Synchronous Crossbars

While some components of the system will be designed specifically for the project, commercial components will be used whenever suitable parts are available. We have investigated a wide range of commercially available components and have considered various options for semi-custom components, including GaAs gate array technology.

For the opto-electronic components and transmission coding circuits, one of our objectives is to use integrated multi-channel components. Ideally, we would like to have single components capable of handling 32 channels. While commercial components have not yet achieved this level of integration, acceptable multichannel components are now becoming available. The most attractive optoelectronic modules now available are produced by Siemens and can support 12 channels with over 1 Gb/s per channel. These devices use VCSEL technology to integrate multiple laser diodes on a single substrate. While these parts are capable of only limited transmission distances (300 meters), they are suitable for local area applications and can certainly support a very effective demonstration of burst switching technology. Vitesse is now advertising a four channel electronic transceiver circuit that implements the Fiber Channel transmission standard. These parts incorporate both line coding and clock recovery functions. The combination of the Siemens and Vitesse parts allow all the transmission interface functions of a 32 channel burst switch port to be implemented with just 14 components (vs. 96 with alternative single channel components).

The data path portion of the burst switch requires the implementation of a large crossbar (256×256). We have evaluated a variety of components, including a 64×32 GaAs part, capable of 1 Gb/s data rates. Using this part to build a larger crossbar requires fanin and fanout components. These can be implemented using GaAs gate arrays but the limited number of I/O pins available in GaAs gate arrays means that only 4 fanin and fanout circuits can be implemented in a single part. This leads to a crossbar containing 32 crossbar components and 64 fanin/fanout components. Moreover, our analysis of the timing and synchronization requirements has led us to conclude that an asynchronous crossbar will not provide a satisfactory solution, in any case. What is needed is a

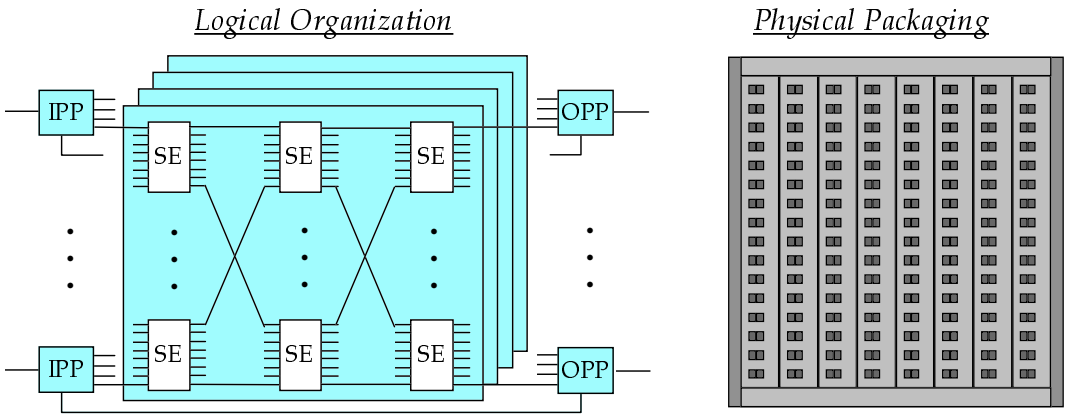

Figure 6: 160 Gb/s ATM Switch

synchronous crossbar in which the inputs include phase adjustment circuits to align the received data to the local clock. The two main candidates for implementing the required crossbar components are GaAs gate arrays and custom CMOS circuits. While CMOS circuits are more limited in data rate, they can support much higher gate complexities and much higher I/O pin counts. This makes them a viable alternative if parallelism can be used to overcome the more limited data rates. Our analysis of the alternatives shows that the required 256×256 crossbar can be implemented with just 16 CMOS chips operating at 135 MHz, but would require at least 128 components using GaAs gate arrays operating at 1.2 GHz (including both crossbar and fanin/fanout parts). The use of lower speed signals does increase the number of I/O pins required on printed circuit boards significantly. However, with appropriate physical partitioning, the number of signals that must pass through I/O connectors on any single board can be kept under 1,000, which is within the capability of modern high density connector technology.

4. 160 Gb/s ATM Switch

Any new communication technology must operate within the context of a variety of existing networks. Burst switching is no exception. To demonstrate the viability of burst switching in a realistic network context, burst switches must be capable of interfacing directly to existing network components, such as ATM switches or IP routers. To support an effective demonstration of burst switching, we are implementing a 160 Gb/s ATM switch through which end users will be able to send and receive data using the prototype burst switch. This switch is being built using components developed in previous research projects. All of these components are also being used in the control portion of the burst switch. One of the three ATM switch components (the Switch Element) is being modified to provide capabilities needed within the burst switch.

Figure 6 shows the logical organization and physical packaging of the planned configuration. The system is built around three key components, the *Input Port Processors* (IPP) which perform the

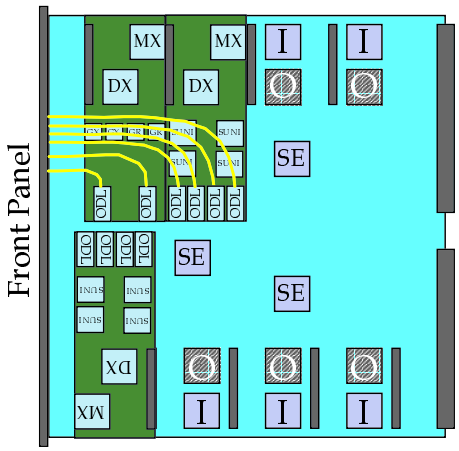

Figure 7: ATM I/O Module

ATM virtual circuit lookup at the input port of the switch, the *Output Port Processors* (OPP) which queue cells awaiting transmission on the outgoing links and the *Switch Elements* (SE) which are arranged in a multistage interconnection network, and deliver cells to the proper output (or outputs) using the routing information supplied by the IPPs. The system will be packaged in a single shelf (approximately 60 cm tall) in a standard equipment rack.

Most of the components of the system will be packaged within a set of eight I/O Modules illustrated in Figure 7. Each I/O module contains eight IPPs, eight OPPs and the associated transmission interface circuitry. In addition, each I/O module includes a total of eight SE chips, implementing one first stage switch element and one third stage switch element. The IPPs, OPPs and SEs will be mounted on the I/O Module's main printed circuit board, and the transmission circuitry will be mounted on eight line cards connected to the main board through a set of high density connectors. Two types of line cards are currently planned. One will carry four OC-12 interfaces and the other will carry a pair of serial interfaces based on the Hewlett Packard G-link transmission components. However, it appears that new SONET components that are now becoming available may make it possible to implement an OC-48C line card instead of the G-link line cards. Work on the G-link line cards is being deferred to allow a fuller evaluation of this possibility.

Each I/O module has a set of high density connectors which plug into a passive midplane, as shown in Figure 8. A set of additional printed circuit boards carrying the center stage switch elements will be mounted on the reverse side of the midplane. The midplane will provide the necessary interconnections to link the first and third stage switch elements with the center stage switch elements.

5. Plans

During the next quarter, we plan to continue to develop details of the burst switch architecture, including developing a detailed understanding of the control issues assocoiated with the burst setup

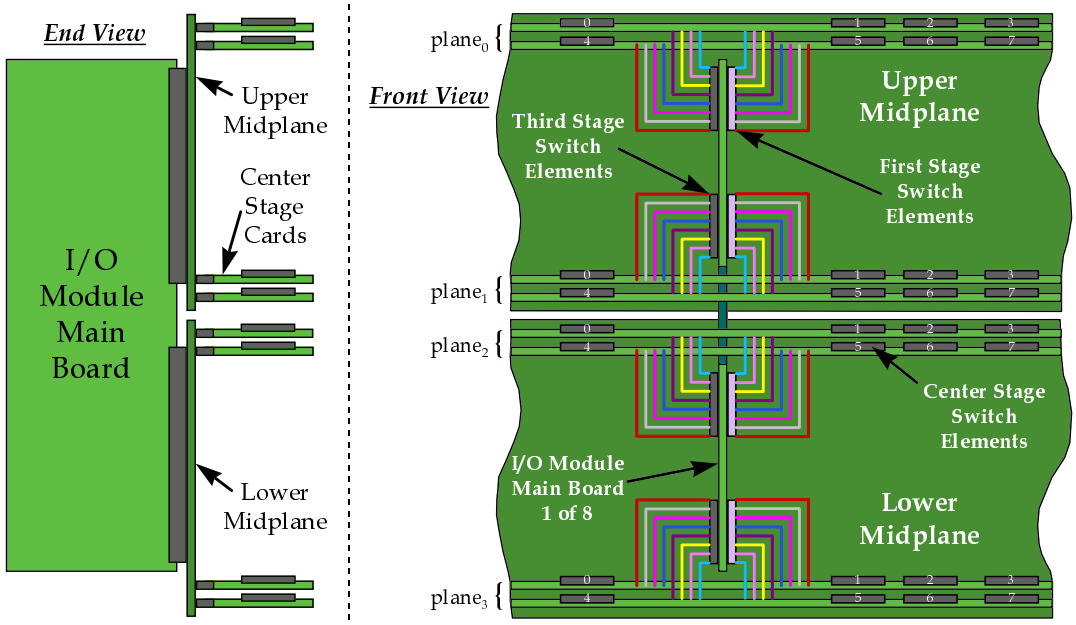

Figure 8: Interconnection of ATM Switch Components

process. We will also continue the engineering efforts on the ATM switch and will address the interconnection of ATM and burst switches.

References

[1] Chaney, Tom, J. Andrew Fingerhut, Margaret Flucke and Jonathan Turner. "Design of a Gigabit ATM Switch," *Proceedings of Infocom*, April 1997.

[2] Turner, Jonathan S., "An Optimal Nonblocking Multicast Virtual Circuit Switch," *Proceedings of Infocom*, June 1994.

[3] Turner, Jonathan S. "Terabit Burst Switching," Washington University Technical Report, WUCS-97-49, 1997.