Terabit Burst Switching

Jonathan S. Turner
jst@cs.wustl.edu

WUCS-97-49

December 18, 1997

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

Abstract

This report summarizes the results of an architectural study on *Terabit Burst Switching*. The purpose of this study was to explore alternative architectures for very high performance switching for data communication, using a combination of optical and electronic technologies. We explore two alternative implementations of the burst switching concept in detail, one using a hybrid architecture with an electronic core, and an integrated architecture using an all optical data path. We also briefly discuss an approach using optical TDM. Our results show that using the hybrid architecture, it is feasible to construct systems with aggregate capacities of tens of terabits per second and that efficiently support highly dynamic data communication applications with individual channel rates of 10 Gb/s. The integrated architectures allow greater scalability and superior performance, but their viability rests on the development of inexpensive optical wavelength conversion devices. Appropriate devices could be available in five to ten years.

[0]This work was supported by the Advanced Research Projects Agency and Rome Laboratory.

Contents

List of Figures

Terabit Burst Switching

Jonathan S. Turner
jst@cs.wustl.edu

1. Introduction

Communications technology has made remarkable strides in performance over the last decade or so, with the development of economical fiber optic media, gigabit transmission and multiplexing electronics and high performance ATM switches. While we have seen impressive progress, the fact remains that current applications of fiber optic transmission systems use only a small fraction (less than 0.1%) of the available bandwidth in the fiber and the only real potential for using more capacity is through point-to-point multiplexing. A variety of all-optical or "mostly-optical" network designs have been proposed to more fully exploit the tremendous potential of optical technology, but in general these proposed architectures have lacked the flexibility needed to effectively support bursty data applications operating from low speeds to very high speeds, and have lacked the ability to incrementally evolve to higher performance as technology continues to advance. This report examines architectural alternatives for *Terabit Burst Switching*. We seek to develop system designs with the following characteristics.

- Effective support for both continuous and bursty applications, with peak rates up to 2.4–10 Gb/s in near-term implementations.
- Effective statistical multiplexing of bursty data applications, allowing average link utilizations of 75% with highly bursty traffic.
- Evolvability, so that as electronics and optical technology continue to improve, the number of assignable channels on a fiber can grow and the data rate of individual channels can increase.
- Minimal logic and data storage requirements in high speed data channels, allowing for incorporation of high speed technologies for which logic and storage costs are high.
- Transparent accommodation of timing variability in the network, precluding the need for strict synchronization of network components with extremely high precision.
- Interoperability with conventional ATM technology, so that users connected via inexpensive ATM network interfaces can still exploit the burst network.

Burst switching is not an entirely new concept. The ideas presented here are similar to fast circuit switching concepts developed in the early eighties [1, 2]. Differences arise from the systematic inclusion of buffering, multicast and datagram-style routing of bursts. However, the principle reason to reconsider burst switching now is that limitations on electronic processing speeds, makes it unlikely that conventional packet or cell switches will be able to support link rates much above 10 Gb/s, at any time in the future. Fiber capacities however, reach well into the Tb/s range. To exploit this capacity effectively for data communication, a different approach is required. Burst switching has the potential to make the full bandwidth of fiber optic transmission systems directly accessible to dynamic data applications in a scalable and cost-effective way.

In Section 2 of this report, we introduce the overall concept of terabit burst switching, describe a generic architecture and identify the key performance and cost issues. In Sections 3 through 5, we study three specific architectures for terabit burst switching, two based on WDM transmission links and one based on TDM.

2. System Concept

Optical communication technology has remarkable capabilities with respect to high bandwidth, long-distance transmission. At the same time, optical device technology has very limited capabilities for implementing logic and memory. Terabit burst switching seeks to provide effective support for dynamic data communication using a hybrid approach that combines the strengths of optical and electronic technologies.

2.1. Basic Operational Principles

Figure 1 shows the basic concept for a terabit burst network. The transmission links in the system carry multiple *channels*, any one of which can be dynamically assigned to a user data burst. Network access can be via point-to-point links or using *access rings* as shown in the figure. Either wavelength division multiplexing (WDM) or time division multiplexing (TDM) can be used to provide multiple channels. Whichever multiplexing technique is used, one channel on each link is designated a *control channel*, and is used to control dynamic allocation of the remaining channels to user data bursts.[1] The switching systems can be implemented in various ways. *Hybrid architectures* include a separate control subsystem to process control information and forward it to the proper outgoing links. *Integrated architectures* tie the data and control components more closely together. Hybrid architectures can take advantage of existing technology components and consequently offer implementation advantages in the short term. Integrated architectures offer greater scalability and are likely to be more cost-effective in the long term. In this report, we focus initially on hybrid architectures, but a fully integrated architecture is discussed in Section 4.

[1] In addition to WDM and TDM, there is a third alternative, which is not to multiplex channels on a single fiber, but to provide a separate fiber for each channel in a multi-fiber cable. This can be more cost-effective in applications where the transmission distances are relatively short. In fact, even in a wide area network that uses WDM or TDM on inter-switch trunks, access links may be more cost-effectively implemented using parallel fibers.

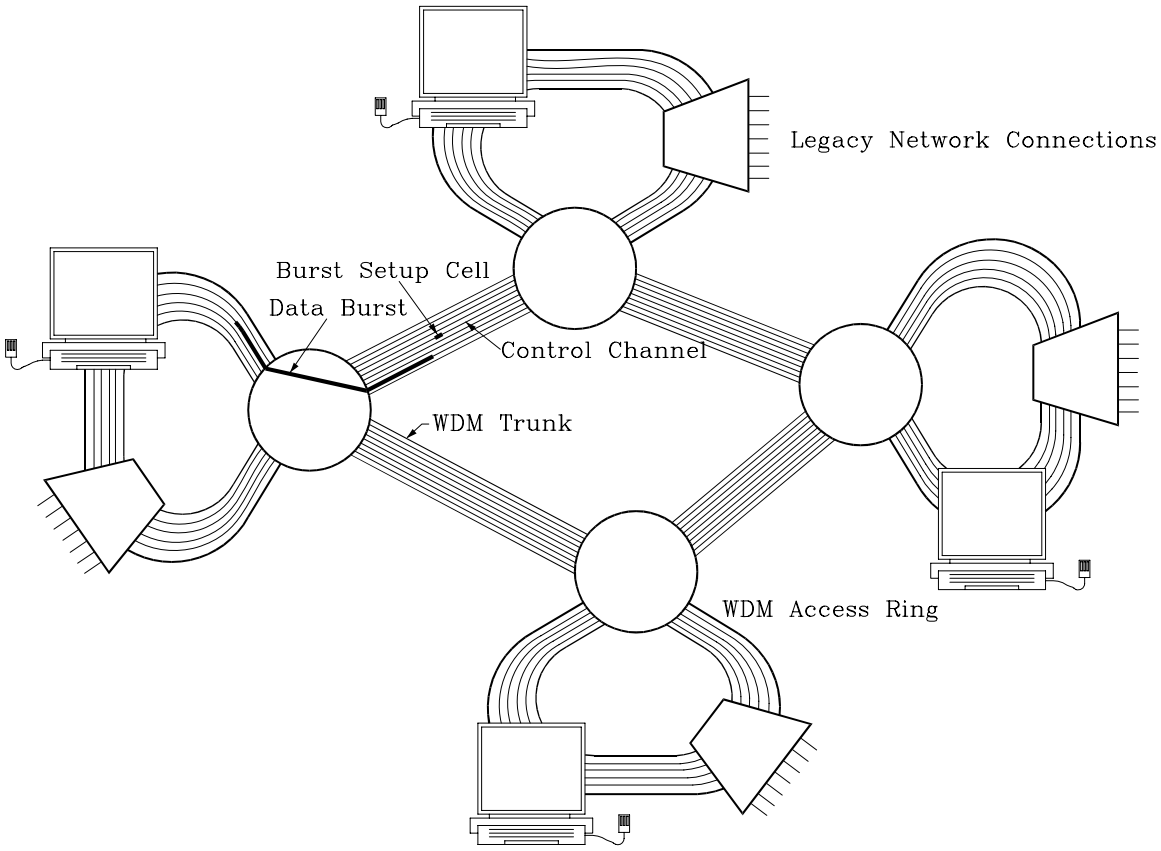

Figure 1: Terabit Burst Network Concept

The hybrid architectures considered here use a conventional ATM switch in the control section of the burst switch. The other major component is a high speed circuit switch, capable of connecting individual input channels to output channels. The control channels on each fiber are switched through the circuit switch to ports on the ATM switch. User terminals can be connected either to the ATM switch or directly to the circuit switch, depending on their performance requirements. Those connected to the ATM switch have just a single WDM or TDM channel.

When a user, who is connected to a circuit switched port on the burst switch, has a burst of data to send, an idle channel on the access link is selected, and the data burst is sent on that idle channel. Just before the burst transmission begins, a *Burst Setup Cell* is sent on the control channel, specifying the channel on which the burst is being transmitted and the destination of the burst. The switch, on receiving the setup cell, selects an outgoing link leading toward the desired destination with an idle channel available, and then establishes a path between the specified channel on the access link and the channel selected to carry the burst. It also forwards the setup cell on the control channel of the selected link, after modifying the cell to specify the channel on which the burst is being forwarded. This process is repeated at every switch along the path to the destination. When the burst transmission is completed, a *Burst Tear-down Cell* is sent by the user, and the switches along the path

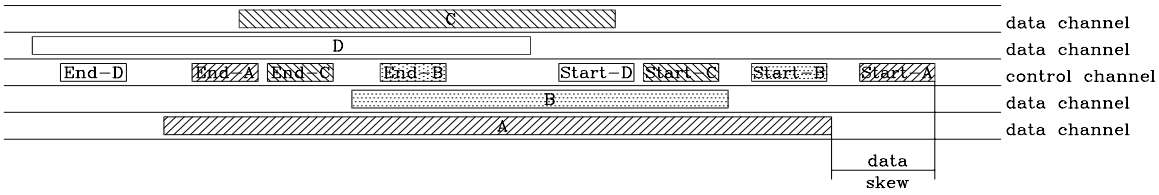

Figure 2: Burst Transmission Timing

use this to release the channels on the links and remove the connections through the circuit switches. To avoid lost network capacity, due to occasional loss of tear-down cells, users are required to send setup cells repeatedly during a burst, in order to hold the channel open. The switches automatically tear down channels for which no setup cell has been received within a timeout period.

Figure 2 shows an example of how setup and tear-down cells (labeled 'Start' and 'End' in the figure) proceed down a link in advance of bursts traveling on parallel data channels. The control cells are sent a short time ahead of the bursts, to give the switches at each hop time to configure connections before the burst arrives. To minimize overhead in the usage of the links, this time skew between the control cells and the data should be kept as short as possible.

There are two alternative ways to perform the burst setup process. In the *Datagram mode*, setup cells include the network address of the destination terminal, and each switch selects an outgoing link dynamically from among the set of links to that destination address. This requires that the switch consult a routing database at the start of each burst, to enable it to make the appropriate routing decisions. In the *Virtual Circuit mode*, burst transmissions must be preceded by an end-to-end route selection process, similar to an ATM virtual circuit establishment. During route selection, the forwarding information associated with a given end-to-end session is stored in the switches along the path. Setup cells include a Burst Virtual Circuit Identifier (BVCI), which the switches use in much the same way that ATM switches use an ATM VCI. Note that while the route selection fixes the sequence of links used by bursts in a given end-to-end session, it does not dictate the individual channels used by bursts. Indeed, channels are assigned only while bursts are being sent.

2.2. Generic Switch Architecture

Figure 3 shows the organization of a generic hybrid burst switch. The circuit switch at the top, terminates the multi-channel links, while the ATM switch at bottom terminates single channel links. The ATM switch and the circuit switch are connected by a number of single channel links. There must be at least one such link for each multi-channel link connected to the circuit switch, but additional links can be provided for handling either ATM switched communication paths, or for burst transmissions to and from terminals connected only to the ATM switch (this is discussed further, below). The ATM switch includes a connection processor which performs ATM signaling and is responsible for overall system management.

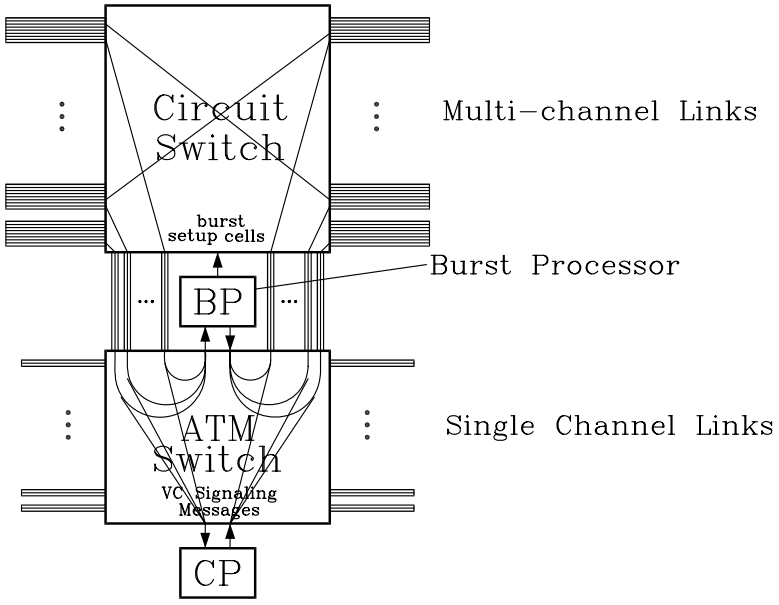

Figure 3: Hybrid Burst Switch

The dynamic assignment of channels to bursts is implemented by a *Burst Processor* (BP), which is connected to one or more ports on the ATM switch and which has a control interface to the circuit switch, allowing it to establish paths through the circuit switch as required. Arriving control cells pass through the circuit switch, along the permanently switched paths to the ATM switch, and from there, pass through the ATM switch to the BP. Similarly, outgoing control cells pass through the ATM switch to the circuit switch, and thence to the outgoing link. The channels used to carry control cells can also be used to carry lower speed user data traffic. These can be switched as ATM virtual circuits within the ATM switch. If more bandwidth is needed for lower speed, ATM-switched channels than can be accommodated on one channel per link, additional channels can be permanently switched through the circuit switch to the ATM switch.

Users connected by an ATM port can send data either using conventional ATM virtual circuits, or using the dynamic burst setup protocol. In the latter case, burst setup cells are sent on the access link, followed by the data, as before. In this case however, since there is only one WDM or TDM channel, the burst is sent using a previously idle ATM *virtual circuit*. The setup cells are sent to the BP, as before, and the BP selects an outgoing link and channel for the burst, connects the outgoing channel to an output port on the ATM switch through the circuit switch, and switches the virtual circuit being used by the data burst through the ATM switch, to the ATM switch output. In a similar way, bursts arriving at a switch can be routed through the ATM switch to a terminal with only an ATM connection.

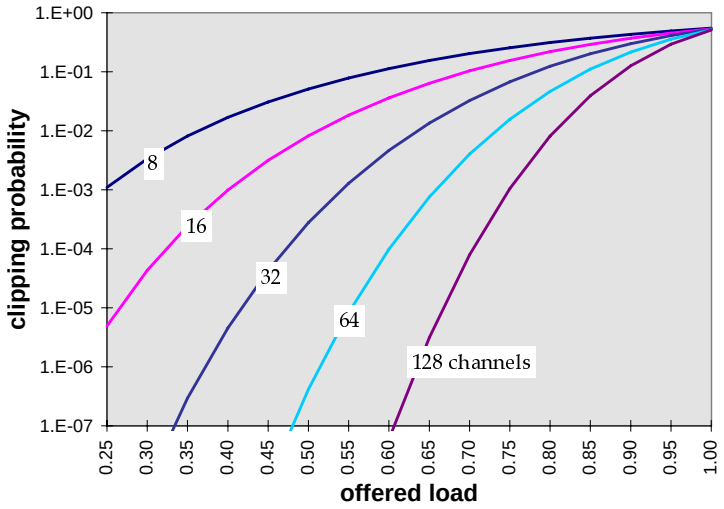

Figure 4: Clipping Probability for Burst Switch

2.3. Performance and Cost Issues

There are two key performance metrics for a burst switch. The first, is the burst clipping rate. Bursts can be clipped when an arriving burst cannot be switched through to its destination, either because all channels to the destination are busy or because there is no available path through the switch. In the case of virtual circuit mode burst switching and a nonblocking circuit switch (any idle input channel can be connected to any idle output channel), the burst clipping probability is easy to estimate. Consider an output link with k assignable channels that is shared among $n > k$ burst virtual circuits, each of which is transmitting data with probability p. The burst clipping probability is the probability that an arriving burst finds that all k channels are in use by the other $n-1$ virtual circuits. If clipped bursts are connected to output channels when an output channel becomes available, the clipping probability is given by

$$\sum_{i \geq k} \binom{n-1}{i} p^i (1-p)^{n-1-i}$$

For large n, the binomial distribution approaches a Poisson distribution and the above expression simplifies to

$$e^{-\rho k} \sum_{i \geq k} \frac{(\rho k)^i}{i!}$$

where ρ is the offered load (that is, the average number of bursts in progress divided by the number of channels). Figure 4 shows burst clipping rates for a range of offered loads and channel counts. Note that burst clipping rates of 10^{-6} or less can be achieved with offered loads greater than 50% when the number of channels per link exceeds 64.

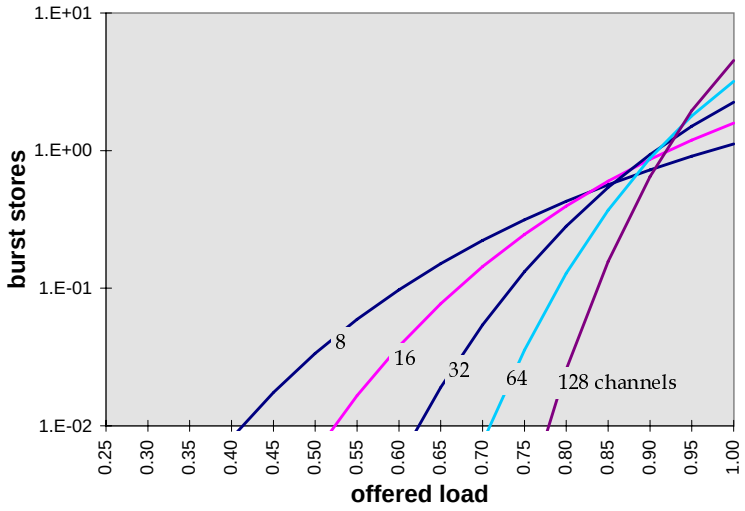

Figure 5: Average Number of Burst Stores Used by One Link

One can reduce the clipping probability by adding memory to the burst switch. In particular, some number of ports of the circuit switch can be used for one or more *burst stores*. If an arriving burst cannot be immediately forwarded, it can be diverted to an unused burst store. Once an output becomes available, the burst can be forwarded through the circuit switch to the selected output link and channel. The average number of burst stores used by bursts arriving for a given output link is given by

$$e^{-\rho k} \sum_{i \geq k} (i - k) \frac{(\rho k)^i}{i!}$$

Figure 5 shows the average number of burst stores used by arriving bursts under a variety of conditions. Notice that for offered loads up to 90%, the average number of burst stores used remains less than one. However, a single link may occasionally need a larger number of burst stores. Figure 6 shows the number of burst stores a link needs to ensure that the probability of not having enough burst stores for arriving bursts is less than 10^{-6}. If burst stores are shared among the output links of a large burst switch, then the total number needed to ensure small clipping probability will be close to the average number required. The results here indicate that a burst switch with n links (for moderately large n) can be operated at link loads close to 90% with low burst clipping rates if the number of burst stores is at least $2n$.

The second key performance metric for burst switches is the overhead associated with assigning a channel. Since the control circuitry requires some time to process and forward burst setup cells, there must be some time skew between the time at which a setup cell is sent and the time the data burst is sent. If the data bursts encounter little or no delay at each switch, then the time skew must be sufficient to cover the processing time at every switch along the path (so that the data burst will not overtake the setup cell). Suppose

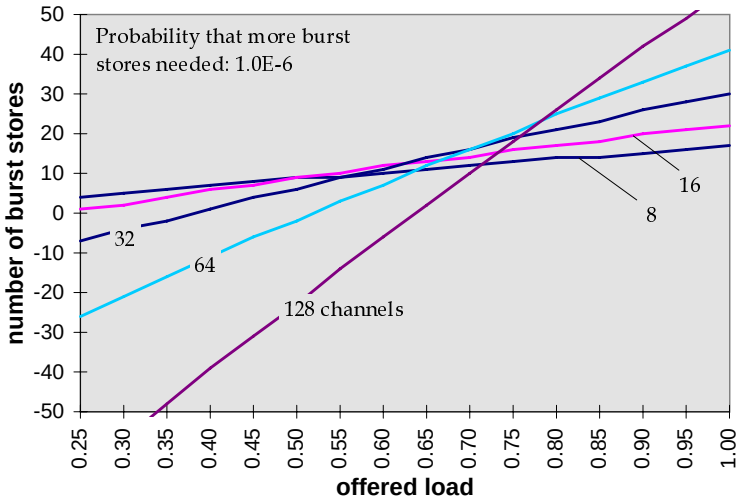

Figure 6: Number of Burst Stores Needed to Ensure Small Probability that Too Few are Available

the setup process can be accomplished in about 20 μs per switch. Then a path containing 10 switches will require a control/data skew of 200 μs. If the channel is used for a period that is at least ten times this delay, then the efficiency of utilization of the channels can be acceptably high. With a 2.4 Gb/s channel rate, this implies burst sizes of 600 Kbytes for high efficiency on ten hop paths.

The required gap between the setup cell and the burst can be greatly reduced if the switches introduce delay in the burst data paths to compensate for the control delay. This approach is illustrated in Figure 7. At every input, the data channels of the multi-channel links are delayed relative to the control channel. For a 20 μs delay on a 2.4 Gb/s channel, this requires either 6,000 bytes of memory per channel or 6 km of fiber optic delay line per link. If a fixed delay is added to the data paths, the required control/data skew is dependent only on the variation in the delay that can be experienced by the control cells. This variation can be reduced to very small values if cells are time-stamped on entry to the ATM switch and the time-stamps used to schedule the forwarding of control cells at the output of the ATM switch. Since it's possible for several cells at the ATM switch output to have the same scheduled transmission time, some deviation from the ideal delay is still possible. If this deviation is encoded in the control cells as they are forwarded along the path to the destination, subsequent switches can adjust the delay that they impose on control cells in order to compensate. This prevents the delay variation from accumulating from switch-to-switch, allowing a small skew to be inserted at the sending end and then maintained throughout. With a skew of 1 μs, high efficiency can be obtained with bursts as small as 3 Kbytes.

Further improvements can be obtained by scheduling the circuit switch control operations to occur just before the bursts arrive, rather than when the setup cells arrive. With

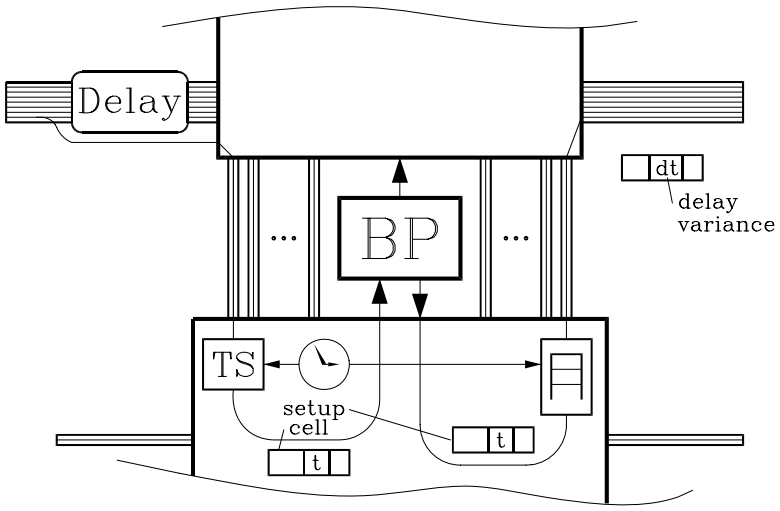

Figure 7: Skew Management

this technique, it should be possible to reduce the switching overhead to 100-200 ns, making the system efficient for burst durations of 1-2 μs. With 2.4 Gb/s channels, this gives efficient operation for bursts as short as 300–600 bytes. For 10 Gb/s links, we obtain efficient operation for bursts of 1200–2400 bytes.

There is another limiting factor on the efficiency of short bursts, since setup and teardown cells must be sent at the start and end of each burst. With cell lengths of 53 bytes, we only get efficient operation at burst lengths of at least 1 KB. In addition, burst lengths are limited by the bandwidth of the control channel relative to the data channels. A single control channel serving h fully loaded data channels can handle average burst lengths no smaller than about $2h$ cells, or roughly $100h$ bytes. A network requiring efficient operation with smaller average burst lengths may require multiple control channels on links with a large number of data channels.

3. Technologies for Implementing Terabit Burst Switching

To implement high performance, yet cost-effective burst switching systems, several component technologies are needed. The availability of these component technologies will ultimately determine the technical and economic feasibility of burst switching, and will shape the architectural choices that must be made.

3.1. Integrated Opto-Electronic and Electronic Device Arrays

Hybrid architectures using WDM transmission convert optical signals to electronic form for switching. This requires a large number of opto-electronic conversions. For example, a system using 64 WDM channels on each of its fibers will require 64 conversions per fiber. If each opto-electronic conversion must be implemented in a discrete component, this

can become prohibitively expensive. Costs can be dramatically reduced using integrated devices. Bellcore has implemented a prototype laser array containing 20 DFB lasers, two semiconductor optical amplifiers (SOA) and a 20×4 star coupler. The device can supply signals on any of eight wavelengths (2 nm spacing) and data rates of 2.5 Gb/s. Hughes hase developed an InP HBT chip with eight integrated photodetectors and preamplifiers. With these levels of integration, eight components are sufficient to provide the opto-electronic conversion needed for a 64 channel fiber.

Temperature control is a significant issue for WDM components, since the wavelengths produced by lasers are strongly dependent on temperature. In telecommunications applications, the problem is exacerbated by the requirement that equipment operate in enviornments where the temperature varies widely. Controlling device temperatures in the face of wide ranging environmental temperatures requires bulky and expensive temperature control components. The impact of this can be reduced somewhat by reducing the number of lasers whose wavelength must be precisely controlled. This is done by generating output signals using external modulation rather than direct modulation. The system provides one precisely controlled laser for each output wavelength used, and the output of this laser is used as a carrier for modulators that are provided for each output fiber. In this way, a system with n output fibers and h channels per fiber uses just h lasers, rather than nh. It does require nh modulators, but these do not require the precise temperature control required for the lasers.

3.2. GaAs Circuit Technology

GaAs circuit technology can be used to provide the required high speed electronic components for hybrid burst switch architectures. These include transmission line coding and receiver components, as well as crossbar switches.

Rockwell [14, 8] produced an eight channel transmitter array using their GaAs HBT technology, that supports transmission speeds of 2.5 Gb/s. A similar, eight channel receiver array, including clock recovery circuits has also been developed. In addition, Rockwell has produced 16×16 crossbar chips capable of operating at 10 Gb/s. Devices with this level of integration are required for practical WDM burst systems based on hybrid architectures.

Lower speed components are available commercially from Vitesse [13]. In particular, they provide a four channel Fiber Channel transmitter and receiver array that operates at 1 Gb/s and implements a 5/4 line code and full clock recovery. It provides an eight bit interface on each of the four channels, allowing lower speed circuits to interface to it directly. Vitesse also provides a 64×64 crossbar supporting data rates of 200 Mb/s. Using byte wide data paths, this device could support burst switching at individual channel rates of over 1 Gb/s.

3.3. Wavelength Conversion

Burst switches using WDM require wavelength conversion [10] devices. There are two main categories: *fixed converters* convert from a single fixed input wavelength to a single fixed

output wavelength and *tunable converters* allow either the input wavelength or the output wavelength to be variable.

For some burst switch architectures, fixed converters are sufficient. The most practical approach currently is to simply convert the optical input signal to electronic form using a PIN diode and then converting it back to optical form using a laser with the required output wavelength. This also allows for electronic timing regeneration, to improve signal quality. The drawback is that it does not maintain the signal in optical format end-to-end, sacrificing optical transparency. All optical wavelength converters are starting to become practical. The most promising approach use optically controlled gates to enable an input signal at one wavelength to modulate a carrier frequency at the desired output wavelength. Typical designs use a pair of semiconductor optical amplifiers in a Mach-Zender interferometer.

Wavelength converters that convert one wavelength from a WDM input signal to a particular output wavelength can be implemented by preceding a fixed converter by a wavelength selector. The wavelength selector is implemented using a grating to distribute the input wavelengths across an array of optical gates (typically clamped gain SOAs), and selecting the desired wavelength by enabling the appropriate gate in the array.

3.4. Optical Space Switch Technologies

Optical space switches are a key component of any burst switch architecture that maintains the data in optical format throughout [5]. There are two principal options for optical space switching. One is based on lithium niobate electro-optic directional couplers. This is fairly mature technology and switches as large as 8×8 have been developed. These switches do introduce significant loss, limiting the number of stages of switching that can be used before amplification is required.

The second principal alternative uses passive waveguides and SOA-based gates. This technology is less mature than lithium niobate, but 4×4 components have been reported. This approach is promising, since SOAs are compact enough to allow higher levels of integration, and because SOAs provide gain, facilitating the construction of multistage switches.

3.5. Optical TDM Components

Burst switches can be constructed using optical TDM instead of WDM. WDM and TDM switches can be constructed using very similar architectures. The principal difference is that where WDM systems employ wavelength conversion, TDM systems employ timeslot conversion. Timeslot conversion requires the ability to select one of a set of input timeslots and then time shift the information in that timeslot prior to multiplexing the signal into an output bit stream. The timeshifting operation can be done by routing the signal through a set of precision delays of exponentially increasing lengths. SOAs or some similar optical gating components are needed to select which delays a given signal passes through. The selection of a given input timeslot can be done using a Terabit Optical Asymmetric Delay (TOAD) device [9].

4. System Design Using WDM Links and Electronic Switching

There is a variety of possible approaches for implementing a terabit burst switch. In this section we consider a system using WDM links and electronic switching. The h wavelengths on each input fiber are demultiplexed and converted to electronic signals, which are then switched through a high bandwidth, electronic circuit switch. On the output side, groups of h outputs from the circuit switch are multiplexed onto h wavelengths on the output links. This approach allows signals on any input link and any input wavelength to be switched to any output link and wavelength. While it's possible for the circuit-switched path to pass analog signals, jitter and cross-talk will limit the achievable data rates too much to make this approach practical in most applications. We therefore assume digital transmission with clock recovery and regeneration at every switch. This issue is discussed in more detail below.

4.1. Reference System Configuration

In this section, we describe a reference design for a terabit burst switch. This will provide a baseline configuration for performance studies and for comparisons to other architectural approaches. The reference design is shown in Figure 8. The system supports two types of multi-channel links, one with 64 channels and one with 8 channels. The 64 channel links are appropriate for connecting switching systems together, while the 8 channel links are appropriate for connecting to terminals, which require much lower average channel utilizations, and can consequently be adequately served by fewer channels. The reference system has 30 of the 64 channel links and 240 of the 8 channel links. The individual channels can be operated at either 2.4 Gb/s or 10 Gb/s. Channels that are connected through to the ATM switch are limited to 2.4 Gb/s.

The core of the system is a three stage Clos-type interconnection network. This network is strictly nonblocking, so that any input (link,channel) pair can be connected to any output (link,channel) pair. The interconnection network has a total of 4096 ports. Of these, 128 are used to connect to the ATM switch, another 128 are used for burst stores and the remainder are divided evenly between the two types of multi-channel links. It is constructed from 64×128 crossbars in the first stage, 64×64 crossbars in the middle stage and 128×64 crossbars in the third stage.[2]

The ATM switch has 512 ports in the core fabric, each capable of supporting a 2.4 Gb/s interface. 128 of these are used to connect to the circuit switch, eight are used to connect to the Burst Processor and one is used to connect to the Control Processor. The remainder are available for supporting user connections.

[2] An alternative design is also possible, using a five stage Clos network. This network could be constructed using smaller crossbars (16×32 in the first and second stages, 16×16 in the center stage and 32×16 in the fourth and fifth stages), making it easier to build the required components. On the other hand, it would be somewhat more complicated to control.

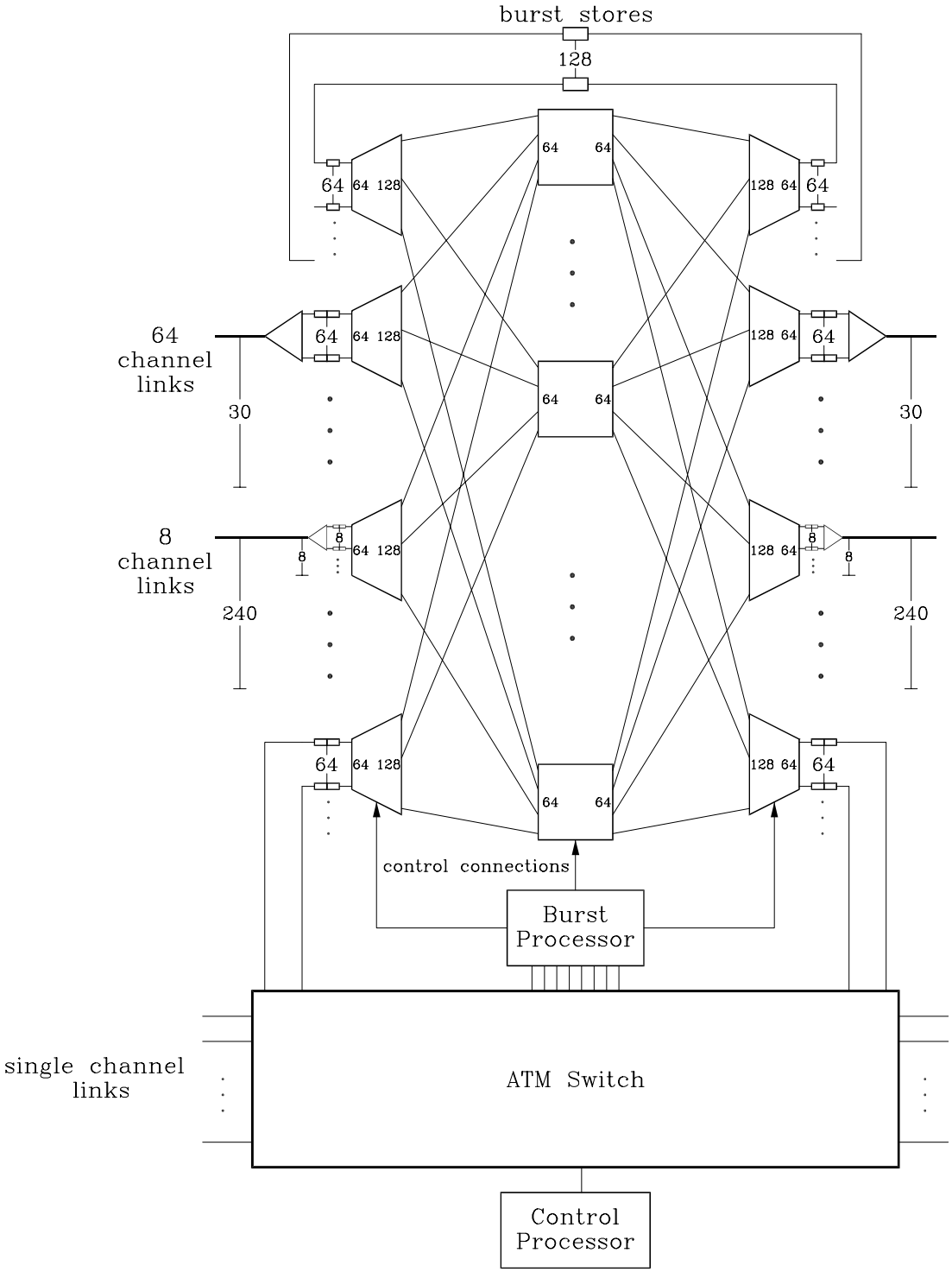

Figure 8: Reference System Using WDM Links and Electronic Switching

4.2. Circuit Switch Components

Ideally, the circuit switch would be implemented using integrated circuits that directly implement the required crossbars. Alternatively, they can be built up from smaller components. For example, the 64×128 crossbar could be constructed using 32 crossbars with 16 inputs and outputs, together with some additional chips to provide fan-in and fan-out. These can be packaged on a single printed circuit board without difficulty. GaAs HBT technology is suitable for implementing these components.

Each of the three columns in the circuit switch, has its own set of control signals. These include a `close_crosspoint` signal, an `open_crosspoint` signal, a `crossbar_number` and an (`input`,`output`) pair. Each crossbar in a column has its own number and must be able to decode that number, and respond appropriately to the control signals. The decoding logic required for this should be able to process a control operation in under 15 ns.

4.3. Transmission Components

The received optical signals are converted to electronic form and re-timed upon input. This involves recovering clock from the received data, then using the clock to sample the data before sending it on through the circuit switch. Data can also be regenerated on the output side of the system, to eliminate jitter that may be introduced by the circuit switch. If necessary, regeneration may also be done at intermediate stages.

To obtain a reasonably economical implementation, multiple transceiver channels should be integrated onto a single chip. Implementing multiple channels on a single chip can be difficult if they must be designed to handle independent data inputs, since in this case, each receiver channel must have its own phase-locked loop for clock recovery, and these circuits are difficult to isolate from one another when they are on the same chip. Failure to adequately isolate the circuits results in coupling of the phase-locked loops, leading to imperfect tracking of the individual data sources and unacceptably high bit error rates.

This problem can be avoided, if instead of simply regenerating received signals using the recovered clock, signals are forwarded using a local clock. Because the local clock can be slightly different in frequency from the recovered clock, some mechanism is needed to make timing adjustments. This can be accomplished by using a transmission format that allows the periodic insertion or deletion of padding bits. For example, if data is sent on the links in the form of ATM cells, a few padding bits can be inserted between every pair of cells. At any point where signals must be re-clocked using a different frequency source, timing adjustments can be made by adding or removing padding bits. If the variation in frequency of different timing sources is limited to .01%, then it's sufficient to have plus/minus one bit of adjustment for every 10,000 bits of data. This suggests that at most three padding bits would be needed between successive ATM cells, in a network that used ATM transmission formats. Note that circuits that perform timing adjustments must be able to identify the padding bit locations in order to make the adjustments. An alternative approach involves the use of a 4B/5B transmission line coding, as used in Fiber Channel transmission components, currently operating at rates of 1 Gb/s. These line codes are "word-oriented"

and include idle words that can be used for making timing adjustments. In a burst network, it's sufficient for a sending terminal to insert at least one idle word for every 10,000 data words sent. Switches in the network can then insert additional idle words or remove them, as necessary to accommodate end-to-end timing variations.

The use of a local clock to forward signals allows multi-channel transceivers to be implemented more easily, because the different channels on a given fiber can all be driven by a common clock at the sending end. This means that a multi-channel receiver need not implement independent clock recovery circuits for each channel. Indeed, it can recover clock from just a single channel and then use this to sample data from all channels. The individual channels still require independent phase adjustments, to compensate for the different propagation speeds associated with the different wavelength channels. However, this can be done without introducing coupling effects. Within the circuit switch, a single local clock can be distributed to all circuits, eliminating the need to do clock recovery internally. While phase adjustments are necessary to compensate for skew, this can be done fairly simply. This approach allows multiple transceiver channels to share a single chip, allowing substantially higher levels of integration. Another benefit of this form of synchronization is that it eliminates timing acquisition delays which occur when a given transceiver's data source is switched. Timing acquisition delays are not a problem if they can be limited to a fraction of the shortest burst length for which efficient operation is needed (say a few hundred nanoseconds). However, achieving such short acquisition delays can be difficult in high speed transmission systems, where the time constants of the clock recovery circuits are usually in the millisecond range, to provide long-term stability.

Using this approach, and assuming that eight optoelectronic devices can be integrated on a single chip, and that an eight channel electronic transceiver can be implemented on a single chip, the receiving section of a 64 channel optical link can be implemented on a single board containing an optical amplifier, the demultiplexing component and 16 additional packages.

4.4. Burst Stores

Each burst store provides the ability to buffer data from a single arriving data burst. Burst stores can be implemented as a circular buffer in a memory, with associated control logic to start and stop data storage and forwarding. For external channels with rates of 10 Gb/s, each burst store must have a memory bandwidth of 20 Gb/s. With 212 bit data paths to and from memory (212 is half the number of bits in an ATM cell), this can be accomplished with SRAMs having a cycle time of 11 ns. Using memory chips having 32 bit data paths, each burst store can be implemented with 7 chips. With chips of 1 Mbit in size, this leads to a burst store size of just under 1 Mbyte. 16 burst stores could be packaged on a single printed circuit board, and this could perhaps be improved to 32.

4.5. Burst Processor

In order to process burst setup cells, the burst processor must perform the following sequence of steps.

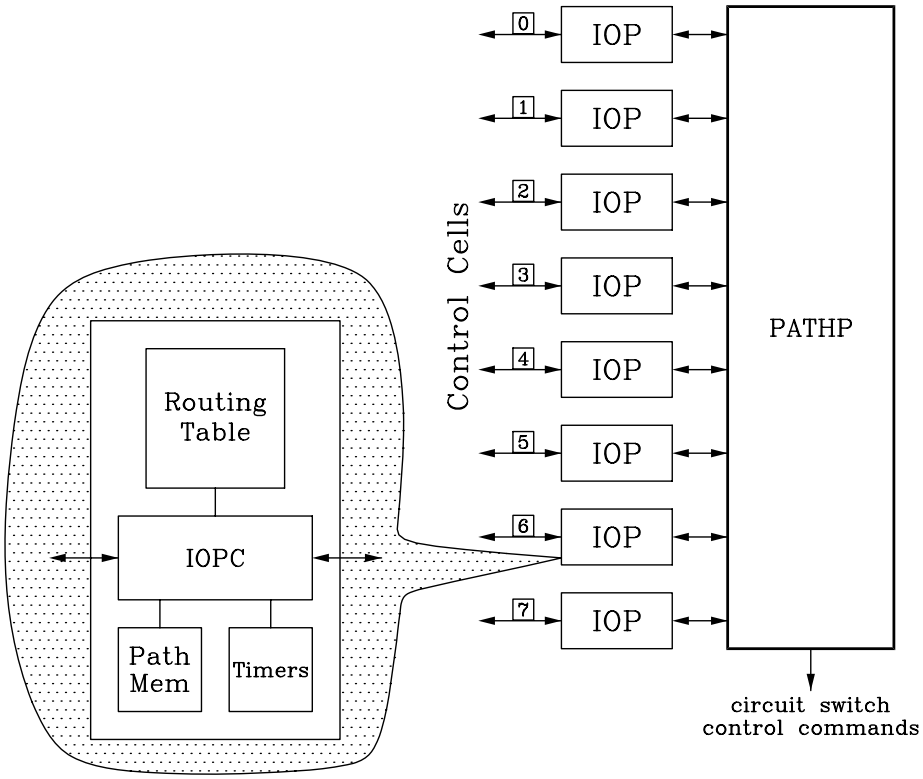

Figure 9: Eight Channel Burst Processor

- *Determine the output link the burst is to be forwarded to.* For virtual circuit mode processing, this requires a table lookup using a BVCI in the setup cell, together with the number of the input link on which the setup cell was received. For datagram mode processing, this requires consultation of a routing database, using a network address in the setup cell.
- *Determine circuit switch ports used by input and output links.* This requires consulting a static table that specifies the range of switch ports used by each link in the system.
- *Select an idle channel on the selected link.* This requires consulting a table which maintains the busy/idle status of every channel.
- *Select a path through the circuit switch.* This requires consulting a table which maintains the busy/idle status of every internal link within the circuit switch.
- *Close the crosspoints that implement the selected path.*
- *Record the path chosen in a table.* This will be used later to tear down the path.
- *Initiate a timer for the path.* This will be used to trigger teardown of the path, in the absence of new burst setup cells.

If no free channel on the desired output link can be found, an idle burst store must be selected to receive the burst. In the absence of any idle burst store, the burst is simply discarded. Similar steps are involved in removing a connection in the circuit switch, either in response to a burst teardown cell, or in response to a timeout.

Because the BP is involved in the handling of every burst, it can easily become a performance bottleneck. To enable high performance systems, with large numbers of channels and short duration bursts, it's necessary use parallelism to increase the burst processing rate. Figure 9 shows a BP design that uses eight-way parallelism to achieve a peak processing rate of about 60 control operations per microsecond. It consists of two types of components. The *IO Processors* interface with the ATM interconnection network, determine the outgoing link to be used for each arriving setup cell, keep track of paths that have been established and manage timers associated with the established paths. Each IOP is responsible for a contiguous range of 512 input ports of the circuit switch. In addition to the IOPs, there is a single *Path Processor* that performs path hunts for the circuit switch and issues control operations to the circuit switch. This design allows one control operation every 64 μs for each of the channels on the multi-channel links. Counting two operations per burst, the system could handle average burst durations of 128 μs. At 10 Gab/s data rates, this implies *average* burst sizes of at least 160 KB.

Figure 9 shows details of the IOP in the inset and Figure 10 shows details of the PATHP. When the IOP receives a setup cell from the ATM switch, it performs a routing lookup to determine the outgoing link the cell should be sent to and then forwards a *path setup request* to the PATHP. This request includes an input link number, an input channel number and an output link number. The PATHP determines what ranges of circuit switch ports the input and output links are connected to, determines if there is an available output channel, and if so finds a path through the circuit switch connecting the input channel to the output. It then sends an acknowledgment to the IOP and sends the appropriate control commands to the circuit switch. The IOP, on receiving the acknowledgement, records the path used by the connection (this information is used at the end of the burst to release the path) and starts a timer to trigger release of the path in the absence of new setup cells required to maintain the path. Since the timer is required only for the case when burst release cells are lost, a fairly large value (tens or even hundreds of milliseconds) can be used, minimizing the impact of timer processing on burst processing capacity. If there is no available channel on the requested output link, the PATHP returns a negative acknowledgement to the IOP. The IOP can then request connection to a burst store, also through the PATHP. The operation is virtually identical to creating a connection to a link. Burst stores are connected to the circuit switch in blocks of 64 consecutive ports, so a single PATHP operation can determine if any burst store in a particular block of 64 is free, and if so, create a connection to it.

The principle challenge involved in the creation of a high performance burst processor is managing the contention for access to the information describing the status of the circuit switch ports and internal links. This is handled within the PATHP by dividing the status information into blocks that can be accessed in parallel. There are three main sections to the PATHP, each of which manages separate pieces of status information (see Figure 10). Path requests come in at the left and pass through three distinct processing stages. At stage `A`, the PATHP determines the range of input ports and output ports used by the

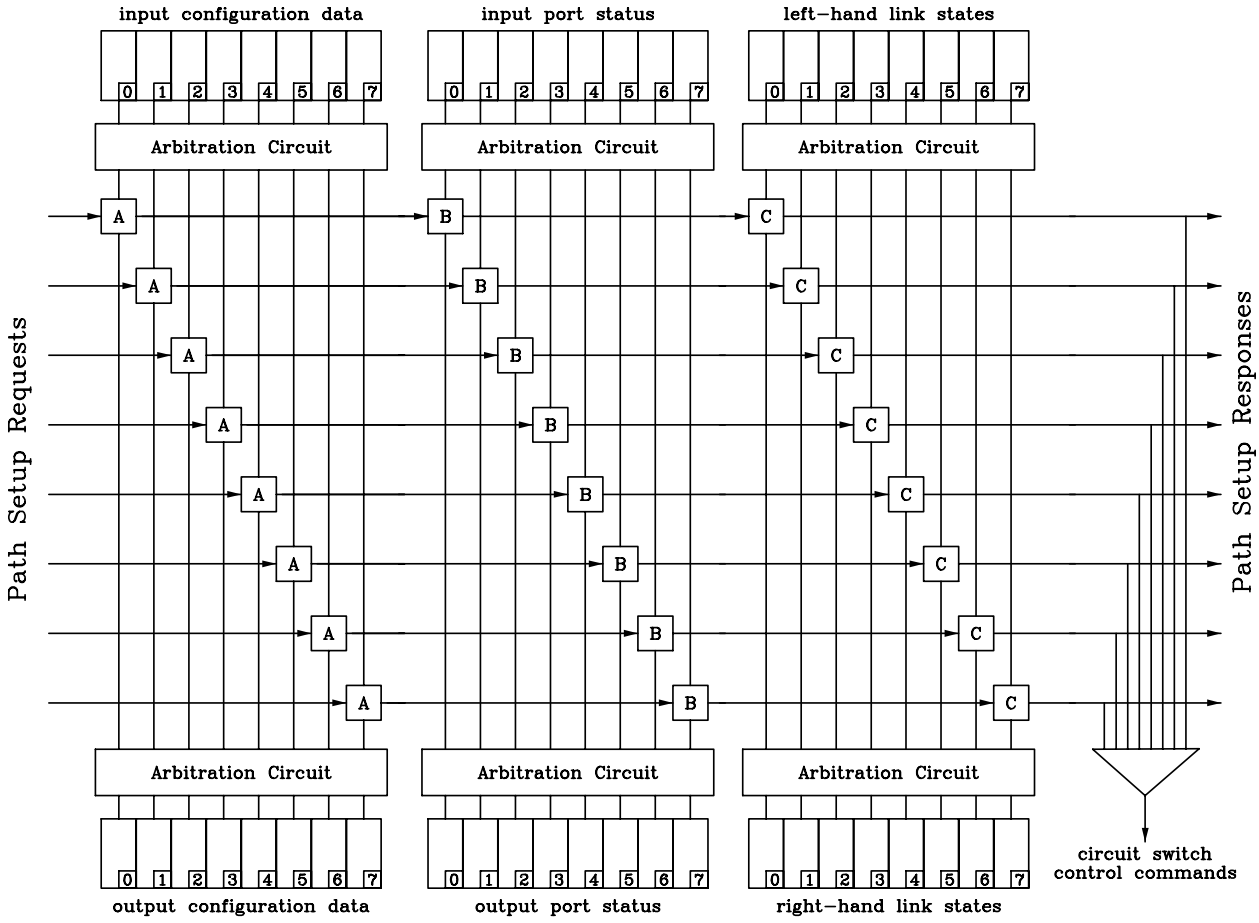

Figure 10: Path Processor

input and output links that are to be connected. In stage `B`, it checks if the input channel on the requested link is idle, and if so makes it busy. Similarly, it determines if there is any idle channel on the requested output link, and if there is, makes it busy. In stage `C`, the PATHP determines if there is any middle stage switch through which it can connect the requested input and output links. In each of these steps, the processing blocks (marked `A, B, C` in the figure) must check some shared status information. Since each of the eight channels may be attempting to access the same status information in parallel, it's necessary to divide the information into pieces that can be accessed in parallel.

The most demanding case is the stage `C` processing. The status information here consists of state bits, specifying the busy/idle status of the links connecting the middle stage switches to the switches in the first and third stages (1 indicates a busy link, 0 an idle link). To enable fast operations, this information is organized into 128 words of 128 bits each. Each word specifies the status of the links connecting to one switch in the first stage, or one switch in the third stage. If we are attempting to connect a first stage switch `x` to a third stage switch `y`, we access the words for `x` and `y`, perform a bit-wise OR operation on the two words and then determine if there are any 0 bits in the resulting word. If there are, then

the middle stage switches corresponding to the bit positions of the 0 bits are accessible from both `x` and `y`. A `C` stage processing block, selects any one of these middle stage switches, sets the corresponding bits in the status words that were accessed and then writes the status words back to the memory. To give each of the processing blocks access to the memory, the accesses passes through an arbitration circuit which connects processing blocks in parallel to the memory bank containing the status bits they require, and holds those connections for long enough for the processing blocks to read the required status information and write the updated status information back. If more than one processing block attempts to access the same memory block, its request is blocked, forcing it to re-attempt on the next cycle. We estimate that four accesses to memory (both read and write) can be completed within the time between successive setup cell arrivals, allowing a system with eight memory banks to keep up with requests from all eight input channels.

4.6. Multicast

Multicast applications are playing a growing role in multimedia and distributed computing. This makes it important to have a strategy for coping effectively with multicast applications. Multicast bursts can be handled similarly to normal data bursts, but some additional time is needed to perform the required setup, since multiple outputs must be connected to a given input. In addition, multicast bursts are more subject to blocking than unicast bursts, since all the outputs that are to receive the multicast burst must have a free channel. The burst stores can play an important role in improving performance for multicast bursts, since they can allow bursts to be forwarded to different outputs as channels become available. In this section, we do not consider the impact of burst stores on multicast burst forwarding, but focus on the simpler case of direct forwarding.

The complexity of the path setup for multicast bursts depends on the strategy used for burst setup. The simplest mechanisms for multicast burst setup increases the probability of burst blocking, but can be effective and economical when burst fanouts are not too large. Consider a multicast burst setup strategy in which multicast bursts are allowed to branch only in the middle and third stage switches. With this strategy, the Burst Processor must find one middle stage switch that has idle links connecting it to each of the required third stage switches and an idle link connecting it to the first stage switch at which the burst arrives. The path processor can be extended to handle connections of this type in a fairly straightforward way. In particular, the `C` stage processing must read status words for each of the third stage switches through which it must connect to reach the desired set of outputs. For a burst with fanout f, we estimate that the time to perform the setup will increase by $(f-1)/4$ times the ATM switch's basic cell processing time (about 130 ns in the system of reference [12]). Thus, the added latency for a connection with fanout 16 is under 2 μs.

The blocking performance for this simple multicast burst setup strategy is strongly dependent on the average load on the external links (p), and on the number of third stage switches through which the burst must pass (f). For the reference switch, the probability that an arriving burst gets to all required third stage switches is approximately

$$\left[1-(1-p/2)^{f+1}\right]^{128}$$

where p is the probability that a channel on the external link is busy. For $p = 1/2$ and $f = 8$, this probability is less than .00005. Note that there is a total of 64 third stage switches, so the maximum possible value of f is 64. Also note that a third stage switch connected to eight channel links can support additional 8-way branching, so even for $f = 8$, there is the potential for supporting total fanouts up to 64. This analysis indicates that the simple strategy is viable for multicast applications with fanouts that are not too large.

It is also possible to make the network nonblocking for multicast bursts. The most efficient way to do this involves handling of multicast bursts in two passes through the space division network, whenever they cannot be handled in one pass, using the simple strategy described above. To implement this approach, the outputs from a subset of the third stage switches are connected back to the inputs of the corresponding first stage switches. An arriving multicast burst that cannot be routed to its desired outputs in one pass is routed from its input to one of these *recycling paths*, and from there is forwarded to all the required outputs, branching only in its second pass through the space division switch.

The total amount of multicast traffic that can be handled with this two pass approach is determined by the number of first and third stage switches that are devoted to recycling connections. If there are m such first and third stage switches, then the reference switch is guaranteed to be able to support up $64m$ WDM output channels using the two pass multicast approach. If 95% of the multicast traffic is handled using the one pass method, then $1280m$ multicast output channels can be handled altogether. For $m = 4$, this allows all of the outgoing channels to be part of multicast connections.

4.7. Multi-Channel Assignment

It is possible that some applications will require more than one channel for the transmission of a high speed data burst. This can be handled by allowing multiple channel numbers to be specified in the setup cell. The simplest way to do this is with a bit vector, having a number of bits equal to the number of channels on the external link.

A multi-channel burst will require more time to setup than a single channel burst. For a burst with h channels, we estimate that the additional setup time will be $(h-1)/4$ times the basic cell time of the ATM switch, or 260 ns for a burst requiring eight channels.

Multichannel bursts are also more subject to blocking than single channel bursts. The blocking probability can be analyzed in much the same way as in the single channel case. In particular, if all bursts require j channels, then the blocking probability is equivalent to what would be experienced in a system with single channel bursts but in which the number of channels per link is reduced by a factor of j. This means that in order to handle multichannel bursts with low blocking probability, we must either limit the traffic on links or provide larger amounts of storage.

4.8. Physical Packaging

In this section, we consider a possible physical packaging arrangement for the reference system discussed above. This packaging appears feasible, assuming moderate levels of de-

vice integration are achieved. In particular, we require single integrated circuits that can implement each of the following functions.

- *Eight channel optoelectronic receiver.* This device takes a single fiber input, separates eight WDM channels and converts the eight optical signals to serial electronic bit streams. It should handle individual channels up to 10 Gb/s.
- *Eight channel optoelectronic transmitter.* This device takes eight electronic inputs (up to 10 Gb/s) and multiplexes them onto a single fiber with eight wavelengths.
- *Eight channel WDM mux.* This component includes a passive optical combiner, multiplexing eight optical inputs to a single optical output.
- *Eight channel WDM demux.* This component includes a passive optical splitter, dividing a single optical input to eight outputs.
- *Eight channel electronic transmitter.* Each of the eight channels in this device transmits data as cells, with from one to three padding bits between successive cells, and scrambling of cell payloads to provide DC balance and frequent enough bit transitions to enable clock recovery at the receiver. An idle cell pattern is sent during transmission slots where no data is available. A one cell buffer is required for each channel.
- *Eight channel electronic receiver.* One of the eight receivers in this device recovers a clock signal from its incoming bit stream and delivers that clock to the other seven. All eight channels do phase alignment of their respective bit streams and must identify ATM cell boundaries, using the HEC field from the cell header. Each arriving bit stream is synchronized to the local timing reference through a four cell fifo.
- *Crossbar chip.* This chip implements either a 16×32 crossbar, a 32×16 crossbar or a pair of 16×16 crossbars. Two configuration pins on the chip determine how it is used.
- *Fanin/fanout chip.* This device contains eight one-to-four fanout circuits and eight four-to-one multiplexors.

Using the above devices, we can construct the following set of modules.

- *Trunk interface module.* This module interfaces to a single 64 channel fiber pair (input and output). It consists of four circuit boards. The first demultiplexes the arriving optical signal into 64 electronic signals, synchronized to the switch's timing reference. This board requires one eight channel WDM demux, eight optoelectronic receiver chips and eight electronic receivers (each of these chips handles eight channels, yielding 64 channels altogether). The 64 electronic channels are sent to the second circuit board which contains a 64×128 crossbar, constructed from 16 crossbar chips, plus 16 fanin/fanout chips. The third board contains a similarly constructed 128×64 crossbar and the fourth board contains eight electronic transmitter chips, eight optoelectronic transmitter chips and a WDM mux.

		Unit Cost			System Cost		
	number	Low	Medium	High	Low	Medium	High
eight channel optoelectronic receiver	496	400	1,000	4,000	198,400	496,000	1,984,000
eight channel optoelectronic transmitter	496	400	1,000	4,000	198,400	496,000	1,984,000
eight channel WDM mux	30	200	1,000	4,000	6,000	30,000	120,000
eight channel WDM demux	30	100	300	500	3,000	9,000	15,000
eight channel electronic receiver	496	200	500	1,500	99,200	248,000	744,000
eight channel electronic transmitter	496	200	500	1,500	99,200	248,000	744,000
crossbar chip	3072	200	500	1,500	614,400	1,536,000	4,608,000
fanin/fanout chip	3072	150	350	1,000	460,800	1,075,200	3,072,000
burst store	128	250	500	1,000	32,000	64,000	128,000
trunk interface module	30	21,100	52,500	172,500	633,000	1,575,000	5,175,000
line interface module	32	20,800	51,200	168,000	665,600	1,638,400	5,376,000
burst store module	2	27,200	59,200	144,000	54,400	118,400	288,000
network boards	128	2,800	6,800	20,000	358,400	870,400	2,560,000
Total System Cost					1,711,400	4,202,200	13,399,000
Cost per Gb/s of Capacity					46	114	362

Figure 11: Cost Analysis

- *Line interface module.* This module is similar to the trunk interface module, but connects to eight fiber pairs, each carrying eight WDM channels. It has the same components as the trunk interface module, except that it omits the initial WDM demux and the final WDM mux.
- *Burst store module.* This module include four boards, one containing a 64×128 crossbar, one containing a 128×64 crossbar and two boards that each contain 32 burst stores.
- *Network boards.* Each of these circuit boards implements a 64×64 crossbar and includes eight crossbar chips and eight fanin/fanout chips.

The entire reference system can be built using 30 trunk interface modules, 32 line interface modules, 2 burst store modules and 128 network boards. This is a total of 384 circuit boards, which can be packaged in five standard equipment racks. If the ATM switch of [12] is used and is configured for 512 links, it requires an additional five equipment racks. The WDM part of the system would have a total throughput of approximately 40 Tb/s, assuming 10 Gb/s per channel, and the ATM part would have a total throughput of approximately 1.2 Tb/s, assuming 2.4 Gb/s per link.

4.9. Cost Analysis

Figure 11 gives a cost analysis for the reference system under a range of different assumptions. (Since most of the key components are not commercially available, the actual numbers should be viewed as only educated guesses. The analysis is useful primarily for identifying the components that contribute to the overall system cost and how their numbers affect the total.) The top section of the table shows estimated the costs associated with individual components, the next section shows the cost of various modules and totals are at the bottom

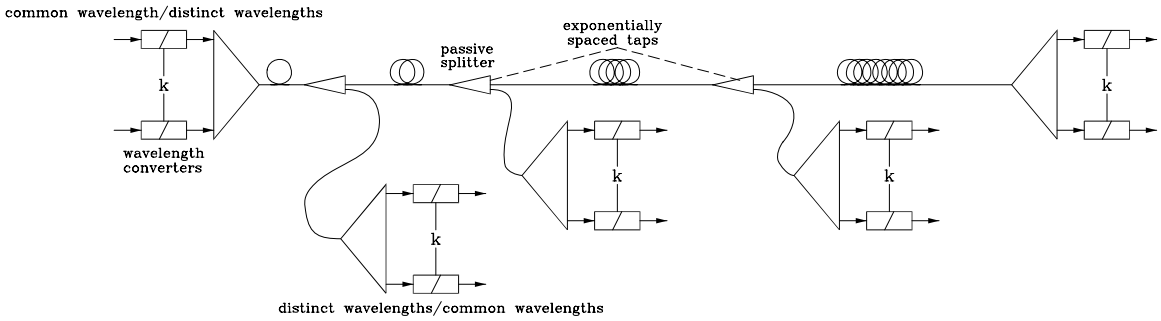

Figure 12: Optical Burst Store

right. The cost of the various modules includes the costs of the components in the modules plus an additional \$100, \$200 or \$400 per circuit board, under the low, mid and high cost estimates.

The overall estimated system cost ranges from about \$1.7 million to \$13.4 million, depending on the assumed component costs, yielding a cost per Gb/s of between \$46 and \$362. For comparison, the system of [12] can be implemented for approximately \$350 per Gb/s of capacity, using 1.2 Gb/s links.[3] So, even under fairly pessimistic component cost assumptions, the system can provide competitive cost/performance, as well as high end performance and total capacity. Using the low cost estimates, the reference system is about a factor of seven better than the system of [12].

It is also interesting to note how the crossbar and fanin/fanout chips dominate the overall component count and consequently contribute significantly to the overall cost. Using the five stage interconnection network mentioned earlier, in place of the three stage network, the number of crossbar chips can be reduced by 1/3. Higher levels of integration can also be applied to the fanin/fanout components, reducing their number and cutting overall system cost. However, these optimizations, while worthwhile, do not yield dramatic changes in the overall cost profile.

5. System Design Using WDM Links and Switching

In the previous section, we considered a burst switching system, using electronic space-division switching. The use of electronic switching components introduces certain disadvantages. First, the timing jitter introduced by electronic components makes it necessary to re-time the signals one or more times within the system. This, in turn means that the data channels must operate at prescribed data rates. If, on the other hand, a communication channel can be switched entirely in the optical domain from end-to-end without re-timing,

[3] In this comparison, 1.2 Gb/s links are used, since there are inexpensive, integrated transmission components at 1.2 Gb/s, while there are no comparable components currently at 2.4 Gb/s. If currently available OC-48 SONET transmission components were used, the cost per Gb/s would increase by \$500 to \$1,000. Since these costs are unrelated to the actual switching costs, it seems inappropriate to make the comparison on this basis.

the only constraints on the information transmitted are those imposed by the optical bandwidth and noise characteristics of the channel. Data can be sent in any format that the endpoints agree on. Indeed, it is possible to send information in analog form, rather than requiring it be digital. Of course, the provision of high quality, wide-band, end-to-end optical channels is a non-trivial proposition. Fibers, optical amplifiers, connectors, splices and optical switching components, all degrade the quality of transmitted signals, limiting the effective bandwidth of end-to-end channels. Designers of terminal equipment must be prepared to deal with worst-case conditions and network operators must maintain the overall infrastructure so that all provided connections meet certain minimum standards of quality. These issues are well-known from experience with analog telephone networks and will not necessarily be any easier to cope with in optical systems. Nonetheless, the possibility of end-to-end optical communication is an attractive goal and it is worth investigating how it might be achieved. In this section, we look at options for implementing the burst switching concept with optical, rather than electronic switching.

5.1. Hybrid System Using Fixed Wavelength Conversion

The reference system of Figure 8 can be converted to an all optical design, by replacing various components, while leaving the overall organization intact. The opto-electronic and electronic receivers in the I/O modules are replaced with wavelength converters that convert the input wavelength to a single common wavelength for use within the switch. Similarly, optoelectronic and electronic transmitters are replaced with wavelength converters that convert from the single wavelength used within the switch to the wavelengths used on the external links. The electronic space division switching components are replaced with optical space division switches of the same dimensions.

Each group of 64 burst stores from the previous design is replaced with a single WDM fiber optic delay line with multiple taps at exponentially spaced intervals, as illustrated in Figure 12. The tap spacing doubles between consecutive taps. This allows a wide range of buffer delays to be handled with a minimal number of taps, while bounding the excess delay imposed by the limited access capability of delay line memories. In particular, the tap spacing guarantees that the delay between the time that an outgoing channel becomes available and the time that the burst is forwarded on the channel is at most equal to the time between the burst arrival and the time of the channel becoming available. Note that electronic burst stores cannot be used here, without compromising the objective of optical transparency. However, for channels that conform to a standard data format, electronic burst stores could be used and may offer cost and performance advantages in those cases.

This system will have similar performance to the system of the previous section. It will also have similar cost characteristics. The key differences will come from the differences in cost between the optical crossbar components and the corresponding electronic components.

5.2. Integrated System Using Variable Wavelength Conversion

Optical technology introduces the possibility of more integrated and more scalable switch architectures. Using these approaches, systems with aggregate throughputs of tens or even

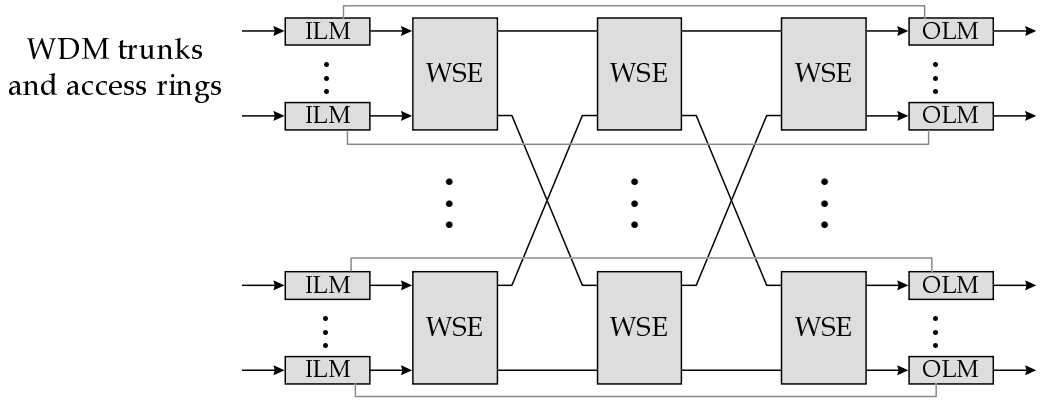

Figure 13: Integrated Burst Switch Design

hundreds of terabits per second may become practical in the five to ten year time frame.

5.2.1. Structure and Operation. Figure 13 is a block diagram of an integrated switch architecture. It comprises three type of components. *Input Link Modules* (ILM) process arriving control cells, determine what network outputs arriving bursts should be sent to, add this information to the control cells and then forward the control cells to a multistage interconnection network. The ILMs also time-stamp the arriving control cells to allow OLMs to time their transmission to the outgoing link.

The interconnection network is made up of *WDM Switch Elements* (WSE), which use the routing information provided by the ILMs to route the cells to the required outputs. In the three stage configuration shown in the figure, the first stage switch elements route bursts across their outputs with the objective of balancing the load evenly across all switch elements in the second stage. From each second stage switch element, there is a unique path to each output, which can be determined directly from the routing information provided by the ILMs. The interconnection network topology is a Beneš network and can be scaled up to larger sizes by adding additional columns of switches to the left and right. A three stage system with eight port switch elements can support 64 external links, while a five stage system can support 512 links.

The *Output Line Modules* (OLM) forward control cells to the outgoing links, using the time stamp information provided by the ILMs to adjust the transmission timing, so as to minimize skew. Each OLM is also connected to its corresponding ILM for control purposes. This control data path allows a remote computer system to manage the routing tables in the ILMs through the use of remote control messages that arrive on the system's external links and are relayed through the interconnection network. This control path can also be used to implement scalable multicast switching. This will be discussed further, below.

Figure 14 shows an input and output line module pair. On the input side, the control channel is separated from the data channels, converted to electronic form and the cells

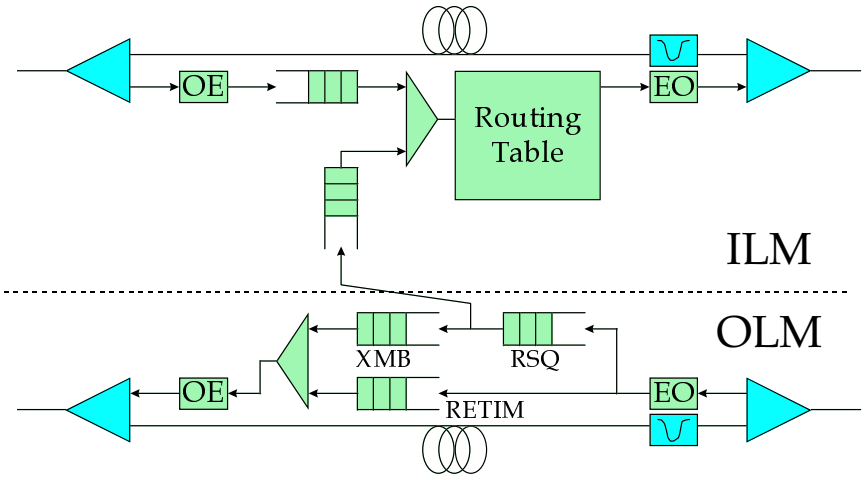

Figure 14: Input and Output Line Modules

are then processed using a routing table. The control cell is then multiplexed back onto the fiber, along with the data channels. The data channels are delayed to match the delay required for the control processing. On the output side, ordinary ATM data cells are passed through a *Resequencing Buffer* (RSQ) and are then passed to a Transmit Buffer (XMB). The resequencer puts cells back into the order in which they entered the interconnection network. The burst control cells are forwarded directly to a retiming buffer which uses the timestamp information added by the ILM to resynchronize the cells with the outgoing data bursts.

The switch element is shown in Figure 15. The figure shows a switch element with eight inputs and outputs and configured for 64 wavelength channels. The switch element consists of three major types of components. The control unit converts the control channels of input links to electronic form, switches arriving cells through an *ATM Switch Element* (ASE) and performs the required control operations to switch arriving bursts. The *Storage Unit* buffers arriving bursts that cannot be immediately switched through to the desired output and the *Data Path Units* (DPU) switch arriving data bursts to the proper output.

Each DPU is associated with one of the switch element's outputs. The data from all eight input fibers are connected to each DPU. An optical space division switch can connect any one of the eight fibers to any of the 64 *Wavelength Converters* (WC) that follow the switch (in particular, the switch can connect an input fiber to multiple WCs). The wavelength converters each have a fixed output wavelength, but can select and convert any one of the input wavelength channels. After passing through the WCs, the converted channels are multiplexed onto the outgoing fiber.

When a burst setup cell is received at the control unit, the routing information in the cell header is used to deliver the cell to the proper output of the ASE. There, the cell is handled by an *Output Controller* (OCTL) for that output. The output controller selects an unused wavelength on the outgoing fiber and issues control signals so that the fiber on

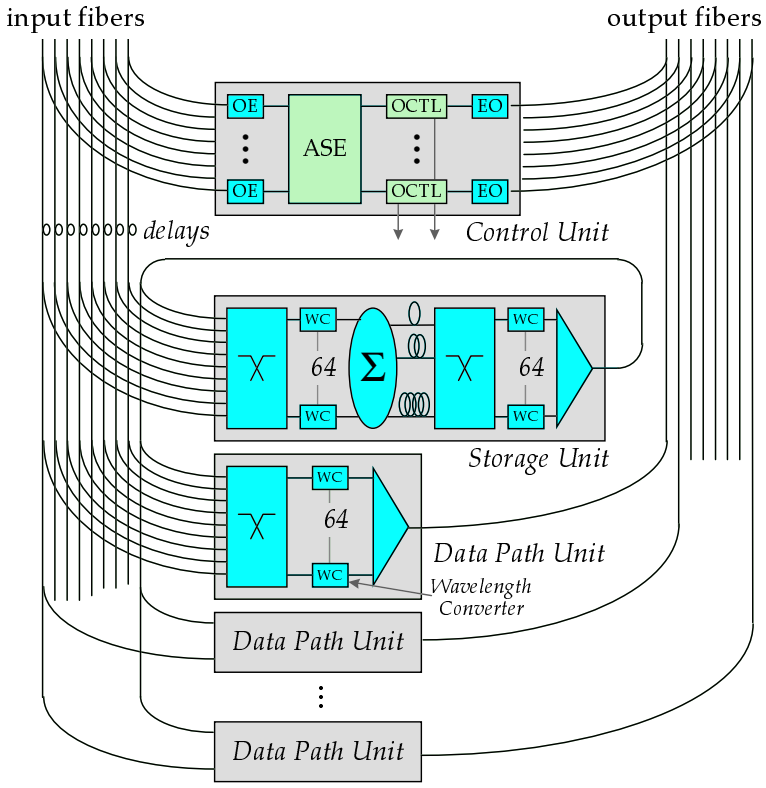

Figure 15: WDM Switch Element

which the burst is arriving, is connected through the optical space division switch of that controller's data path unit. It also issues control signals to cause the wavelength converter for the selected output wavelength to select the appropriate input wavelength.

If a burst arrives at a switch element, when there is no outgoing wavelength for it to use on the required output link, the ASE will divert the link directly to a controller for the storage unit (not shown in the figure). The storage unit can store bursts in fiber delay lines, using a separate wavelength for each burst stored. Any burst diverted to the storage unit is first converted to a wavelength not currently being used by any other stored burst, using exactly the same mechanisms as the data path units. Each stored burst is then broadcast in parallel to each of a set of optical delay lines of exponentially increasing length. When an output channel becomes available on the output link the burst is addressed to, the storage unit connects the appropriate delay line through to a free output channel and the burst is then passed on to the appropriate DPU. The control for this is implemented by having the controller for the storage unit forward the burst setup cell back to an extra input port on the ASE (not shown), and from there to the proper output controller.

5.2.2. Multicast and Multichannel Bursts. The integrated architecture can handle multicast bursts in one of several ways. For multicasts with small fanout, the ILMs can insert the addresses of the required outputs directly in the control cells as they are forwarded to the interconnection network, and the WSEs can use this information to replicate and forward multicast bursts in much the same way that they handle normal data bursts. There is some additional complexity required in the switch elements, particularly for bursts that must be simultaneously forwarded to some outputs, and stored for later forwarding to other outputs. However, these issues can be handled in fairly straightforward ways.

For multicasts with larger fanouts however, the amount of information needed to specify the output port set can become too large. Large fanouts can be handled effectively using a *burst recycling* technique, similar to the cell recycling technique used for multicast in [12]. To implement burst recycling, a subset of the system's output links are connected back to the corresponding input links. Multicast bursts are forwarded to their ultimate outputs through a *multicast forwarding tree* implemented using multiple passes through the switching network. Each node of the forwarding tree has a limited branching factor, corresponding to the practical limit on the fanout in one pass through the system's interconnection network, but arbitrarily large fanouts can be achieved using multiple passes through the forwarding tree. The number of ports that must be dedicated to recycling is a function of the fraction of the system's capacity that is used by multicast applications, but is independent of the size of the system. This leads to a scalable multicast implementation that is cost-effective in both small and large system configurations.

For applications that require more bandwidth than one wavelength can provide, bursts can be forwarded using multiple channels. Multichannel bursts can be implemented even when the network provides direct support only for single channel bursts. Terminals can implement multichannel bursts by using multiple channels of their access links for launching parts of a multichannel burst in parallel. In this case, the switches in the network treat the different parts of the multichannel burst independently and may route them along different paths to the destination. While this adds complexity at the receiving terminal, it allows the network to more effectively balance the traffic load it is required to handle. Multichannel bursts can also be handled directly. For the integrated system design, this adds complexity to the switch elements and leads to an increase in the burst clipping probability. These effects can be reduced somewhat by allowing the components of a multichannel burst to take different paths through the interconnection network and reassembling them at the outgoing OLM. This may require delaying some components of the burst at the OLM to account for different delays experienced in the interconnection network.

5.2.3. Scalability. The integrated system design has substantial scalability advantages over the hybrid designs considered earlier. First, by extending the number of stages of the interconnection network, the overall system capacity can be extended to tens or hundreds of terabits per second. The number of stages grows as the logarithm of the number of ports, giving the system optimal scaling characteristics, as the number of ports grows.

A key aspect of the integrated design's scalability is that it scales up the burst control capacity along with the system's overall data bandwidth. While the eight channel burst

processor for the hybrid design considered in the last section can support high burst processing rates, it is a potential bottleneck in systems with short average burst lengths. The integrated system can handle short data bursts more easily. In fact, the integrated system can accommodate additional parallelism in the control processing to further improve its ability to handle short data bursts and to support more wavelengths per fiber.

5.3. Is Wavelength Conversion Really Necessary?

Wavelength conversion remains a fairly immature and expensive technology leading many researchers to seek to avoid wavelength conversion as much as possible. Wavelength conversion can often be avoided if the selection of the wavelength is done through an end-to-end coordination process. In a network where connections are established through an end-to-end signaling process and then maintained for long periods of time, this approach is feasible and can use network bandwidth almost as efficiently as a system in which wavelength conversion is available at every switch.[4]

In the burst switching context however, the use of wavelength conversion appears to be hard to avoid. The main reason for this is that wavelength selection must be done using local information at the time a burst arrives at a switch. There is no effective way to select a wavelength at a sending terminal, when the burst is first transmitted. The sending terminal does not, and cannot, know what wavelengths will be available at the various switches along the path to the destination. If wavelengths are chosen without this knowledge, the chance that a burst will reach its destination without colliding with another burst becomes small as the traffic in the network increases. The chances can be improved if multiple signals can be forwarded at the same wavelength from a switch. This can be accomplished by replacing each link with multiple parallel links joining the same pairs of switching systems. Two concurrent bursts addressed to a common next hop switch, that arrive on the same wavelength are simply forwarded on different links of a parallel set. However, to match the collision probability of a system with 64 wavelengths per link and full wavelength conversion, we would need 64 parallel links, requiring either that trunks have much larger total capacities or that we sacrifice the ability to carry many wavelengths on one fiber. Neither alternative is attractive in long distance applications.

While wavelength conversion appears to be an essential component for burst switching, a network may have many applications that can be well-served by relatively static connections that can be coordinated on an end-to-end basis. This raises the possibility of a system in which a subset of the links on each channel are dynamically switched, while the rest are configured for longer periods of time using an end-to-end wavelength assignment process. Such systems are certainly possible and may offer the best overall economic trade-off in an environment with a mix of dynamic and static traffic. However, since the focus of this study is on how to handle bursty data traffic, we do not consider this class of systems in any detail.

[4] It should be noted that such systems require that terminals have tunable transmitters and receivers and creates the possibility that an "out-of-tune" transmitter may interfere with transmissions from other users, either intentionally or unintentionally. This can be avoided with tunable filters at switch inputs connected to terminals, but this adds expense and operational complexity.

5.4. Reference System Configuration

We will take the three stage configuration of Figure 13 as the basis for our reference configuration of the integrated design. If the switch elements have eight ports and links have 64 channels each, we get a system with 64 ports that is directly comparable to the hybrid system presented in the last section.

The input and output line modules of the reference system can be implemented fairly compactly, since the number and complexity of the optical components is limited, and the control functions can be implemented using a pair of custom CMOS integrated circuits of moderate complexity, along with a small number of memory chips to implement the routing table. It appears straightforward to implement an ILM/OLM pair on a single printed circuit board. The switch elements present a bigger challenge however. In particular, the data path units, while conceptually simple are difficult to implement compactly and economically using current technology. While there are several optical technologies that can be used to implement the required space division optical switch, no current devices are directly applicable. The wavelength converters can be implemented using clamped gain silicon optical amplifiers (SOA) to select the desired input wavelength and Mach-Zender interferometers for the final wavelength conversion step.

While current devices do not offer the level of integration needed for burst switching systems, it is useful to explore the implications, should suitable devices become available in the coming years. Given a technology that can realize a single 9 input, 64 output optical crossbar in a single component and one that can implement a complete wavelength converter in a single component, then one could implement a single DPU using about 75 optical components. It should be practical to package such a subsystem on one or two circuit boards. This means that an entire WSE could be packaged in one shelf, in a standard equipment frame, and that the 24 WSEs required for the 64 port system could be packaged in three to six frames, depending on the number of shelves per frame. The ILMs and OLMs can fit comfortably in another frame. To complete the system, one would require 64 laser sources to provide carrier signals with precisely controlled wavelengths. The carrier signal for each wavelength would be distributed to all wavelength converters in the system that output signals at that wavelength. Each wavelength converter modulates its carrier, using a Mach-Zender interferometer, to generate the required output.

5.5. Cost Analysis

Figure **??** gives a cost analysis for the reference integrated system. The cost of the required optical components is very difficult to predict at this point, so there is a wide range on the estimated values. Even with these uncertainties however, there is some useful information that one can glean from the table. Note that the number of wavelength converters is much larger than any other component in the system. These are virtually certain to dominate the system cost under any realistic set of assumptions about underlying component costs. This makes it clear that unless wavelength conversion can be implemented at a cost of at most few hundred dollars, it will be difficult for the integrated all optical design to compete with the hybrid optical/electronic design of the previous section.

Figure 16: Cost Analysis for Integrated Architecture

6. System Design Using TDM Links and Switching

Burst switching systems can be constructed using optical time division multiplexing, rather than wavelength division multiplexing. The basic operating principles are identical. The only differences come from the differences in components used to select a channel and convert it to a different channel. In the case of optical TDM systems, this involves the use of TOAD devices and precision-controlled optical delays, rather than wavelength selection/conversion.

The integrated optical switch of the previous section can be converted to an integrated TDM switch by modifying the data path units and storage units of each switch element. In fact, the only elements that really must change are the wavelength converters, which must be changed to timeslot shifters. That is, they must select a time slot from their input data stream and time shift it to a different position in the outgoing time slot. This operation can be implemented using a single TOAD device and a delay line with a range of one frame time (100 ps, assuming 10 Gb/s channels). This is a somewhat simpler operation than wavelength conversion, making it possible that such devices could be implemented less expensively than wavelength converters. To complete the transformation from a WDM to

a TDM system, the ILMs and OLMs must be modified to demultiplex and remultiplex the control cells using TDM techniques.

7. Conclusions

Terabit burst switching is a promising approach to achieving very high performance data communication networks that have the potential for fully exploiting the vast bandwidth of optical fiber transmission systems. The separation of data and control in burst switching systems allows them to implement fairly complex control mechanisms while keeping the data paths relatively simple.

The limitations of current optical devices make systems with all optical data paths prohibitively expensive, when compared to their electronic counterparts. However, advances that can be expected over the next decade could change that situation. In the long term, the combination of more cost-effective optical components and the need for highly scalable systems, could make systems with all optical data paths more competitive. Either WDM or optical TDM can be used in these systems. The crucial point of comparison will be the relative ease of wavelength conversion, on the one hand with timeslot shifting on the other. If either technology can obtain a substantial advantage in the cost/performance of this operation, it is will almost certainly become the more attractive choice.

While WDM and TDM are perhaps the most interesting multiplexing alternatives for burst switching in long distance applications, there is another alternative that merits consideration when distances are relatively short. This is the use of *space division multiplexing*; that is, the use of multiple fibers in a multi-fiber cable to implement the many parallel channels of a burst switching system. This approach avoids the highest cost optical components and yields lower overall system costs when inter-switch distances are short enough that the fiber itself does not contribute significantly to the overall costs.

References

[1] Amstutz, Stanford R. "Burst Switching — An Introduction," *IEEE Communications Magazine*, 11/83.

[2] Amstutz, Stanford R. "Burst Switching — An Update," *IEEE Communications Magazine*, 9/89.

[3] Chaney, Tom, J. Andrew Fingerhut, Margaret Flucke and Jonathan Turner. "Design of a Gigabit ATM Switch," Washington University Computer Science Department, WUCS-96-07, 1/96.

[4] Gnauck, A. H., A. R. Chaplyvy, R. W. Tkach, J. L. Zyskind, J. W. Sulhoff, A. J. Lucero, Y. Sun, R. M. Jopson, F. Forghieri, R. M. Deroisier, C. Wolf and A. R. McCormick. "One Terabit/s Transmission Experiment," *Optical Fiber Communications Conference* (OFC-96), 1996, post-deadline papers.

[5] Gustavsson, Mats. "Technologies and Application for Space Switching in Multi-Wavelength Networks," In *Photonic Networks*, Giancarlo Prati (editor), Springer Verlag 1997.

[6] Ikegami, Tetsuhiko. "WDM Devices, State of the Art," In *Photonic Networks*, Giancarlo Prati (editor), Springer Verlag 1997.

[7] Masetti, Francesco. "System Functionalities and Architectures in Photonic Packet Switching" In *Photonic Networks*, Giancarlo Prati (editor), Springer Verlag 1997.

[8] Pedrotti, K., et. al. "High Speed Circuits for Optical Networks," In *Photonic Networks*, Giancarlo Prati (editor), Springer Verlag 1997.

[9] Prucnal, Paul. xxx.

[10] Stubkjaer, K. E., et. al. "Wavelength Conversion Technology," In *Photonic Networks*, Giancarlo Prati (editor), Springer Verlag 1997.

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

[12] Turner, Jonathan S. and the ARL and ANG staff. "A Gigabit Local ATM Testbed for Multimedia Applications," Washington University Applied Research Lab ARL-WN-94-11.

[13] Vitesse Semiconductor. Advanced Product Information: Multi-Gigabit Interconnect Chip, VSC7214. Available through `http://www.vitesse.com/products/stdprod.html#1`, 4/30/97.

[14] Keh-Chung Wang, Randall Nubling, Ken Pedrotti, Neng-Haung Sheng, Peter Asbeck, Ken Poulton, John Corcoran, Knud Knudsen, Han-Tzong Yuan and Christopher Chang. "AlGaAs/GaAs HBTs for Analog and Digital Applications," *International Journal of High Speed Electronics and Systems*, 1994, pp. 213–252.