Design of an Optically-Interconnected Multiprocessor

R.D. Chamberlain, M.A. Franklin,
R.R. Krchnavek, and B.H. Baysal

June 1998

Computer and Communications Research Center
Washington University
Campus Box 1115
One Brookings Drive
St. Louis, MO 63130-4899

Design of an Optically-Interconnected Multiprocessor

Roger D. Chamberlain
roger@ccrc.wustl.edu

Mark A. Franklin
jbf@ccrc.wustl.edu

Robert R. Krchnavek
rrk@ee.wustl.edu

Department of Electrical Engineering, Washington University, St. Louis, MO

Burak H. Baysal
bbaysal@ee.siue.edu

Department of Electrical Engineering, Southern Illinois University, Edwardsville, IL

Abstract

This paper presents the design of an optically interconnected multiprocessor. The design is oriented to applications where the performance is bandwidth limited in conventional multiprocessors. The system utilizes board-level polymer waveguides to reduce manufacturing costs. The processor interconnection network, called Gemini, *has a Banyan topology and is composed of dual optical and electronic networks. The optical data paths (circuit switched) are used for passing large data blocks and the matched electrical data paths (packet switched) are used for control of the optical interconnect and for short data messages.*

1 Introduction

To take advantage of optical technology in the context of a multi-microprocessor system, the total design must be reexamined with an understanding of the strengths and weaknesses of optical interconnects. We are currently designing and implementing a message-passing, optically-interconnected multiprocessor and in this paper present the underlying ideas and overall design philosophy. The key ideas are summarized below.

- Focus on Bandwidth Limited Applications: Because of the lack of cost effective, highly integrated optical logic and memory, routing and control of optical interconnection networks (ICNs) must currently still be done electronically. Network latencies are therefore not going to improve significantly over all electronic ICNs. However, optical interconnect bandwidths are very high. Therefore, our focus is on developing a system which is oriented towards applications where the performance is interconnection bandwidth limited.

- Design of a Dual Electronic/Optical Network: By creating the appropriate parallel electrical and optical paths and by carefully staging the transmission of separate control and data messages we can design a multiprocessor ICN combining the best properties taken from the electrical and optical domains. We refer to this dual (optical and electrical) network as the *Gemini* network. In addition to providing a high bandwidth path for long messages, this partitioning provides a lower congestion, low latency path for short messages. This idea is similar in spirit to the wave switching interconnect proposed by Duato et al. [4].

- Design for High Integration and Low Cost: In the longer term, optical interconnects will be successful in the multiprocessor arena only if they are cost effective. This will require that optical switches become physically smaller and that the optical media be manufacturable in a manner similar to the way metal lines are laid down on printed wiring boards (PWBs). We are constructing optical paths directly on boards in the form of polymer waveguides while ensuring that the optical properties of these waveguides are appropriate.

- Matching Memory and Interconnect Bandwidths: If the high bandwidths associated with optical interconnects are to be fully exploited, then a redesign of the standard microprocessor memory architecture and its network interface is required. Otherwise, the memory and processor will become the performance bottleneck. The *Gemini* Network Interface (GNI) has an interleaved memory architecture combined with a dedicated and parameterizable gather/scatter engine that supports fast retrieval/storage of messages to non-contiguous memory locations.

The overall system architecture is shown in Figure 1. Central to the system is the *Gemini* ICN which con-
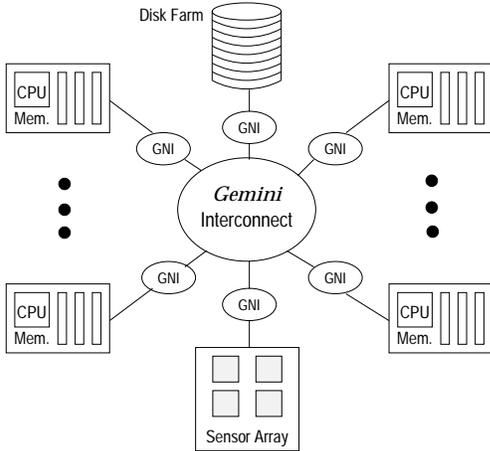
Figure 1: System architecture

nects the various elements through the GNI (*Gemini* Network Interface). The memory systems associated with standard microprocessors are modified to permit dual access by both the processor and the GNI.

2   Applications

Applications whose performance can significantly benefit from the *Gemini* interconnect and the GNI have the following general properties.

Data Bandwidth Is Important: Passing large blocks of data between processors, or between I/O devices and processors, occurs frequently. These data blocks effectively utilize the high bandwidth optical interconnect of *Gemini*.

Computation and Communication Times are Comparable: That portion of the communication which cannot be overlapped with computation is a non-negligible part of the execution time on standard multiprocessors. The *Gemini* system has features that both increase the potential for computation/communication overlap and decrease the time required for communication when they cannot be overlapped.

Short Messages are Present: In addition to long data messages, short messages (e.g., synchronization messages) are present. These messages utilize the lower bandwidth, but less congested and relatively low latency, electrical paths of *Gemini*.

Varying Data Partitioning is Important: The algorithm requires that the data be partitioned and accessed in several different ways (e.g., over the space domain, over the time domain, etc.). The GNI addresses the bandwidth mismatch between conventional memory systems and the optical network during the scatter/gather operation for block decomposed data.

One application class that generally exhibits the above properties are space time adaptive processing (STAP) applications which are common in many defense, medical, scientific, and engineering computing environments. Here, large volumes of sensor-derived data are processed into a smaller data set and presented to a user or automated control system. Critical communications performance bottlenecks occur during data transpose, or "corner turn" operations, and memory access patterns during a transpose are often not sequential [10].

3   The *Gemini* Interconnect

The *Gemini* interconnect consists of a dual network consisting of an optical interconnect for passing large data blocks and a parallel electronic interconnect for both controlling the optical switching elements (and thus message routing) and also for passing small blocks of data. Circuit switching is used for the optical path with an electrical control message setting up the path. The optical path speed is such that using a circuit-switched approach is reasonable. It also avoids problems of providing for controllable optical storage. Messages over the electrical network need not have optical path control functions. They may be short, low-latency data messages which are sent entirely over the electrical network in a self-routing, packet-switched manner.

While many types of interconnection topologies may be employed, to reduce the number of optical switching elements needed, a simple Banyan topology has been selected. While this is a blocking network, it uses only $O(N \log N)$ switching elements, rather than the $O(N^2)$ required for a crossbar topology. In today's optical technology, reducing the number of switches is an important consideration since they are somewhat costly.

3.1   The *Gemini* Switch

Electrooptic 2×2 switching elements are the key devices used in the fabrication of the *Gemini* $N \times N$ optical data path. These switching elements rely on the electrooptic effect (i.e., the application of an electric field changes the refractive index of a material within the field) to provide for pass through and crossover connections between the input and output ports. Thus the state of the 2×2 optical switching element is determined by an electrical control signal. These switching elements can be fabricated using $LiNbO_3$ and are becoming available commercially. Larger 4×4 switches are also available (at higher cost) and we can expect the levels of integration to improve rapidly over the next several years. A typical packaged 2×2 switch in today's technology is about 1.5 cm by 12 cm [9].

Figure 2 shows an $N \times N$ Banyan topology and also illustrates the idea of having dual electronic and optical

networks. That is, in parallel with the optical Banyan network is an electrical Banyan network which is used to control the optical switches and as a lower latency path for short data messages.
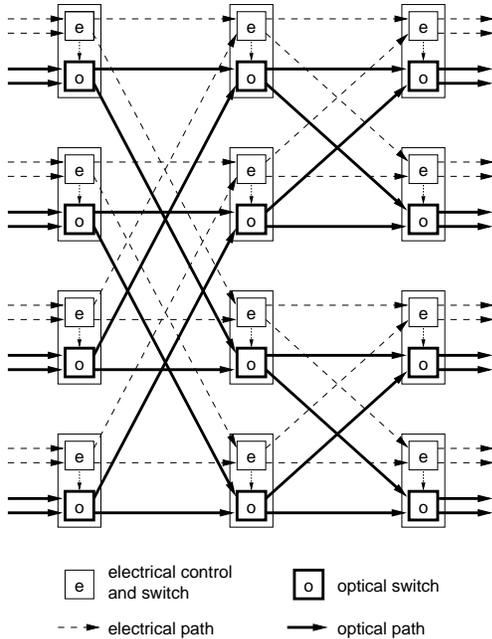

Figure 2: An 8×8 *Gemini* switch

As shown in Figure 3, each processor is functionally connected to the switch in two ways. The first is through an electrical path (via Network Interface Control) and the second is through use of dual-ported memory. This memory is accessible by the gather/scatter engine which, in turn, provides an interface (GNI) between the memory and the *Gemini* network.

3.2   Network Control and Message Passing

Processors communicate with each other by sending messages to and from each other's memories. The general procedure followed is discussed below.

Control Message Transmission: The processor sends a control message through the Network Interface Control and into the electrical portion of *Gemini*. Control messages contain destination information, the memory location for the message, and possibly control information for the gather/scatter engine. t_o is the time when the control message enters the first stage of the network (see Figure 4).

Optical Path Setup: Based on the destination address, the control message is routed through the electrical portion of *Gemini*. At each stage, the electrical switch sets up its paired optical switch (i.e., pass through or cross). t_c is the time required for the con-
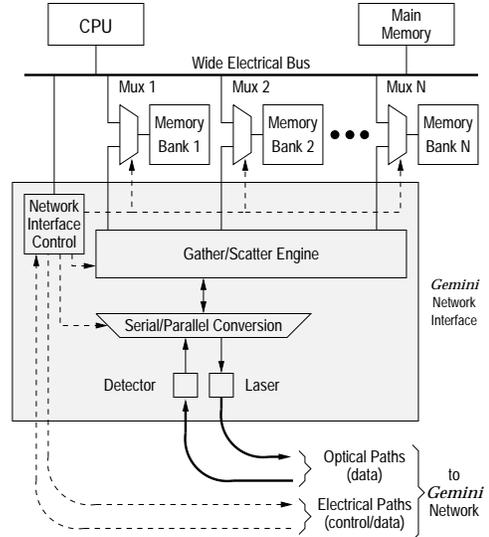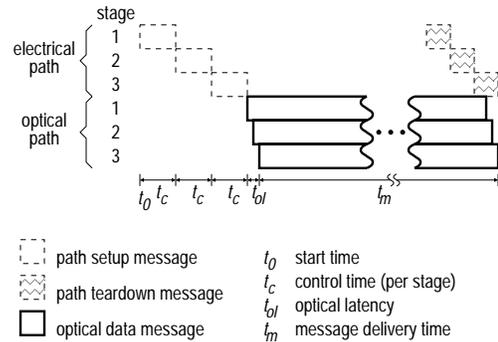

Figure 3: The processor/*Gemini* network interface

Figure 4: *Gemini* network timing

trol message to move from one stage to the next and set up an optical switch.

If at any point the network is blocked, then a return message (i.e., a NACK or Negative ACKnowledgement) is sent to the originating processor telling it to attempt retransmission after a specified length of time (see Figure 5, the NACK message is drawn in a crosshatched manner). The NACK message "releases" the optical switches as it returns to the initiating processor. Note that a partial path release approach which yields slightly higher throughput can be designed for this blocking case. Our initial implementation, however, does complete path release since this is somewhat simpler and avoids any potential deadlock issues. If no blocking is encountered, the control message passes through the entire network in $3t_c$ time (for a 3 stage network), after which a complete optical path has been established.

Data Message Transmission: Given the network

size, the control message length, the electrical path bandwidth, and the setup time needed for each optical switch, a maximum optical path setup time can be calculated. If no NACK message has been received, a timer initiates data transmission at time t_{send} where $t_{send} = t_o + 3 \times t_c$. If blocking occurs, a NACK message will be received by the initiating processor and then, after a fixed delay, a new control message will be sent. The above procedure will then repeat itself. Note that while the return NACK message is in transit, the data block may already be on its way through the network. It will be aborted when the NACK is received, however, a part of it will continue through the optical network until it reaches the blocked switch. At that time it will be routed to the unused (and incorrect) path and be discarded. This is shown in Figure 5 by the short data messages on the optical path.
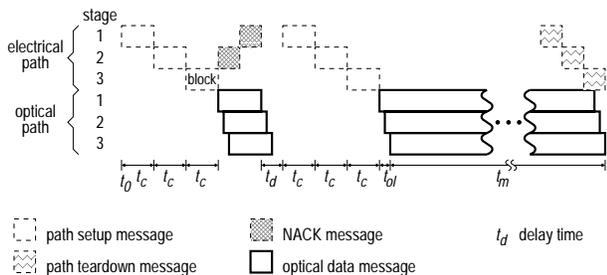

Figure 5: *Gemini* network timing with blocking

Optical Circuit Tear Down: After the data message has completed (or nearly completed) passage through its assigned optical path, the initiating processor will release the optical path which has been held for that message by sending a release control message.

4   Latency Performance of the System

A discrete-event simulation model was developed to determine the performance of the *Gemini* multiprocessor system and to contrast its performance with two alternatives to be explained below. An eight processor system was modeled where setup messages and short data messages (e.g. synchronization messages) are constant in length (16 bytes) while long data messages are exponentially distributed with a mean of 320 bytes. Each of the processors generates messages at the same rate. The message type (e.g., long or short) is a random variable which is set via an input parameter indicating the percentage of each type. Message destinations are uniformly distributed across the processors. The electrical and optical networks have data rates of 800 Mb/s and 20 Gb/s respectively.

Figure 6 below shows average network latency over all messages as a function of message generation rate at each processor. Each curve corresponds to a different percentage of short and long data messages. The general shape of the curves corresponds to queueing system saturation effects. With 100% short messages, which all use the electrical network, performance is good since the network is packet based and there is no need for path setup, teardown and retransmission due to path blocking. As the percentage of long messages increases, these messages, utilizing the optical part of *Gemini*, hold the entire optical path, generate setup and teardown messages, and periodically are blocked after a partial optical path has been setup. This latter event will require attempts at reacquiring the optical path and the probability of this occurring will increase as the percentage of long messages increases.
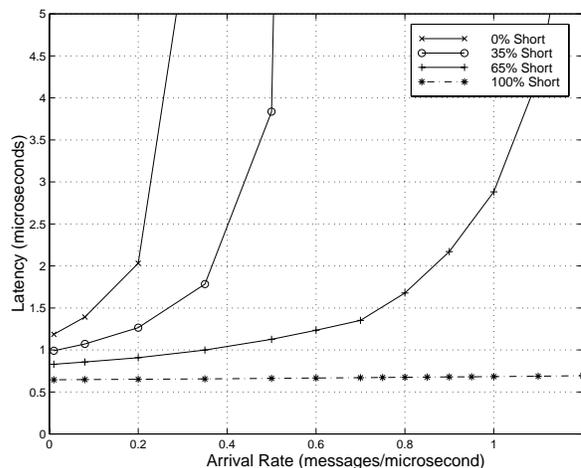

Figure 6: Message latency using the *Gemini* network

Figure 7 demonstrates the value of having a dual network. For this simulation 65% of the messages are short and 35% long. The bottom curve corresponds to using the full capabilities of the *Gemini* network and is also found in Figure 6. The other two curves correspond to two alternative system designs; the first where all data messages (short and long) use only the circuit-switched optical network (setup messages still use the electrical network), and the second where all messages use only the packet-switched electrical network.

The results show dramatically that the dual network significantly outperforms both the all electrical and all optical design options. For the pure optical case, performance is degraded due to several interacting factors. The principal factor is that now corresponding to every message there must be electrical setup and teardown delays. This, in turn, imposes a minimum delay for all messages and also causes the entire corresponding optical path (due to circuit switching) to be held for
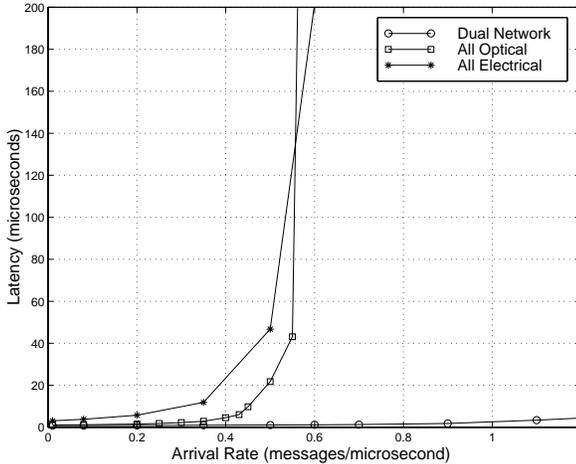
Figure 7: Message latency with alternative networks

Figure 8: Components present in optical path

an unnecessarily long time.

With the pure electrical system the packet properties of the network combined with the fact that most messages are short (65%) results in better performance than the all optical (above a certain arrival rate). Its performance, however, is much worse than the dual network since the long messages cannot take advantage of the high optical path bandwidth.

5   Optical Technology Issues

Electrooptic 2×2 switching elements are the key devices used in the *Gemini* optical data path. These devices are connected together on a single printed wiring board using polymer channel waveguides. Connections between boards use optical fiber.

Figure 8 illustrates the major optical components in an end-to-end optical path. At the source processor, a laser diode is connected to an optical fiber for delivery to the *Gemini* interconnect. The fiber is then coupled into a polymer channel waveguide. The interconnect consists of a number of waveguide bends, crossovers, and optical switches. After the last stage of switching, the waveguide is coupled into an outbound fiber, which is connected to a photodiode associated with the destination processor.

As the optical signal passes through each of these optical components and their interconnections, power is lost. An important design issue is just how large an optical network can be designed before the power levels are too low to be detected by the receiver. While optical amplifiers can be inserted to increase power levels, they are very expensive, take up a good deal of board space, and introduce their own set of distortions which then have to be taken into account. We therefore only
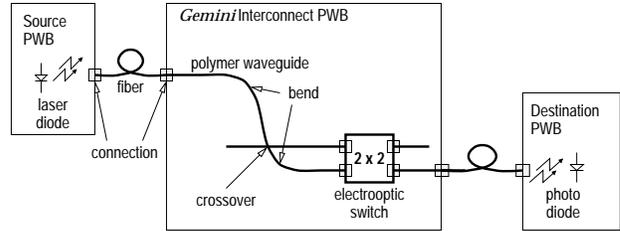
consider designs which do not have such amplifiers.

The component power losses naturally will vary with the technology employed. For the LiNbO$_3$ switching elements and polymer waveguide components used, Table 1 indicates the principal parameters of interest, their approximate current values, and projected values two and four years into the future.

In optical systems, the insertion loss of a device describes the attenuation of the optical signal due to the "insertion" of that device into the path (expressed in dB). Thus, the optical power budget is dependent on the number of connectors and switches in the worst case source to destination path, and on waveguide parameters such as length and number of crossovers and bends in this same path. Losses associated with board-to-board fibers are small enough to be ignored.

We assume that the sources and detectors are positioned with the processors and are separate from the switch fabric. This allows easy replacement of sources and detectors, but necessitates coupling of the source and detector to an optical fiber. The fiber then couples to a waveguide on the switch fabric with this occurring at both the source and detector ends. This results in a total of 4 connector losses. Connector losses associated with switches are included in the switch insertion loss.

We assume the coupling losses for fiber (single-mode) to waveguide (single-mode), fiber to laser, and fiber to photodetector are about the same (0.5 dB) [1]. Currently laser to fiber losses are somewhat higher than 0.5 dB due to beam asymmetry when emerging from the laser, however, the use of Vertical Cavity Surface Emitting Lasers (VCSELs) have reduced such losses. On the other hand, fiber to photodetector losses are currently usually less than 0.5 dB since the detector can be made larger than the fiber core and mode matching is not an issue.

Consider next the LiNbO$_3$ electrooptic 2×2 switching elements. In an $N \times N$ Banyan network, the number of switches in a path is $\log_2 N$. Consider next losses associated with polarization dependence, and losses associated with coupling to the switch. Fortunately, Y-branch LiNbO$_3$ switches of the sort we are using

Parameter	Sym.	1998	2000	2002
Conn. loss (dB/conn.)	P_c	0.5	0.5	0.5
Switch loss (dB/switch)	$P_{2 \times 2}$	5	2.5	1
Crossover loss (dB/cross)	P_x	0.1	0.1	0.1
Waveguide loss (dB/cm)	α	0.05	0.025	0.01
Bend loss (dB/bend)	P_b	0.011	0.011	0.011
System margin (dB)	P_m	4	4	4

Table 1: Device and interconnect parameters

demonstrate a high degree of polarization insensitivity. Such switches have an insertion loss of < 5 dB, a polarization dependence of < 1 dB, and crosstalk of < -35 dB [9, 11].

Regarding the second item, currently available packaged $LiNbO_3$ switches are directly coupled to fiber. We, however, require a coupling to a polymer waveguide and such a coupling will be fabricated as part of our development process. For the power budget analysis we assume that the losses are similar to those measured with switch to fiber couplings.

Since $(\log_2 N)P_{2 \times 2}$ is the insertion loss of all of the switch stages in the longest optical path, with current technology, this is the major loss mechanism in the optical switch fabric. Future technology, however, shows considerable promise toward lowering the insertion loss of the switch elements. For example, high-speed modulators based on polyurethane/disperse red 19 have recently been demonstrated [13]. Further improvements in reducing insertion loss, reducing drive voltages, and increasing device density are anticipated due to several recently synthesized active molecules (e.g., DAST [12]). In addition, semiconductor based electrooptic switches represent an alternative approach to obtaining similar performance improvements. A recent result [14] using InGaAsP-InP quantum wells suggests that 2×2 switch elements could be fabricated with insertion losses as low as 1 dB.

The crossover loss, P_x, is the loss which occurs when two waveguides intersect each other. With single-mode waveguides, losses are minimal (0.1 dB) for intersection angles greater than $30°$. This number is unlikely to change in the future since it is near the theoretical minimum. For the Banyan topology, the maximum number of crossovers in a path is $\sum_{i=1}^{\log_2 N - 1}(2^i - 1)$ for an $N \times N$ network.

The waveguide interconnect medium has an attenuation loss, α, given in dB/cm. While relatively small switch fabrics could be integrated on a single substrate of $LiNbO_3$, and this minimizes interconnect losses, it has two drawbacks as switch fabric size in-

creases. First, the attenuation loss of $LiNbO_3$ waveguides, 0.5 dB/cm, is not as low as some of the current organic polymers used for optical interconnects. For example, photochemically-set, multifunctional acrylate monomers/oligomers from [5] have a reported loss of 0.05 dB/cm at 1550 nm. The second drawback is that there is currently a practical limit to the overall size of the single-crystal $LiNbO_3$ substrate.

As noted above, photochemically-set acrylates have been reported with losses as low as 0.05 dB/cm. Table 1 predicts that waveguide losses will continue to decrease in the next few years. This is due in part because of continued research in materials, but also in part because losses are already lower at certain wavelengths. For example, the fluorinated acrylates described in [5] have losses of 0.03 dB/cm at 1300 nm and 0.001 at 840 nm. Therefore, a change in system wavelength away from 1550 nm will immediately bring about lower loss optical waveguides.

The waveguides interconnecting the 2×2 switch elements consist of the attenuation losses described above plus an additional bend loss. Table 2 shows the longest optical path length, l, as a function of switch size. Using typical physical dimensions for *packaged* 2×2 switch elements, we use a bend diameter of 1.5 cm. The added bend loss is approximately 10^3 dB/km [8] or 0.011 dB additional loss for a $90°$ bend. We have calculated path lengths by assuming a simple layout topology with 1.5 cm separation between switch element inputs/outputs in both the horizontal and vertical direction. Figure 9 shows an idealized layout (not to scale) of the longest path in an 8×8 system. The waveguide bends required are illustrated as $90°$ turns (a conservative assumption). The worst case number of bends in an $N \times N$ network is $2(\log_2 N - 1)$.

Switch Size	Path Length (cm)	Switch Size	Path Length (cm)
4×4	2.2	128×128	98.7
8×8	5.9	256×256	195.4
16×16	12.6	512×512	388.1
32×32	25.3	1024×1024	772.8
64×64	50		

Table 2: Waveguide path lengths

Finally, a system margin, P_m, of approximately 4 dB is required to ensure reliable operation. Summing all of the above elements, the power budget for the worst
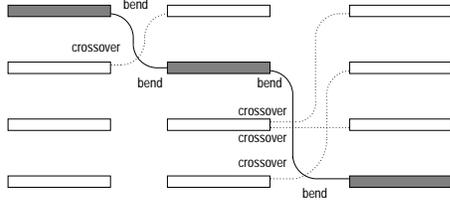
Figure 9: Idealized waveguide path layout

case path is

$$P_T = 4P_c + (\log_2 N)P_{2 \times 2} + \sum_{i=1}^{\log_2 N - 1} (2^i - 1)P_x +$$
$$l\alpha + 2(\log_2 N - 1)P_b + P_m$$

Figure 10 plots the link power budget, P_T, as a function of switch size using 1998 values from Table 1. From the baseline at 0 dB, we subtract the various power loss components in traversing the longest optical path through the switch. The figure presents a graphical picture of where the energy is being dissipated and shows that switch losses are the largest component of the total loss. However, the medium, bend, and crossover losses become a larger proportion of the total link loss as the switch size increases. Finally, the lowest curve in the figure denotes the total link power budget for a given switch size.
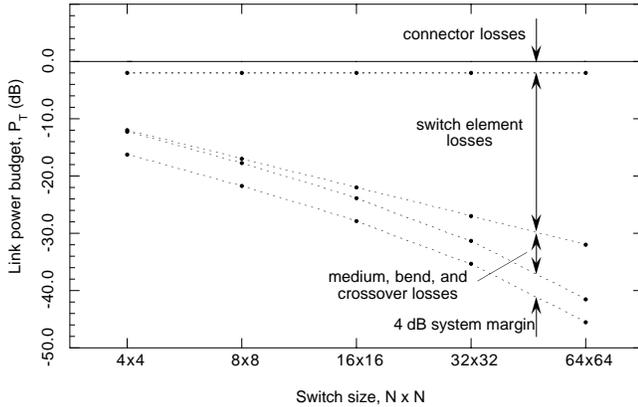

Figure 10: Components of link power budget

Whether a particular link power budget will allow reliable operation of the system is determined by the initial power supplied by the source and the *sensitivity* of the receiver. Receiver sensitivity is the minimum power required to achieve a given bit error rate (BER) for a given data rate. Calculating the sensitivity for a particular photodetector and amplifier requires a detailed analysis of various noise sources; however, recent

receiver designs [6, 7, 15] indicate that a minimum receiver sensitivity of −30 dBm at 10 Gb/s with a BER of 10^{-9} is achievable in the very near future.

Regarding the optical source, high-speed, directly modulated, laser diodes are routinely fabricated with output powers of 1 mW (0 dBm). Considerably higher laser powers are achieved at higher drive-current levels but often require thermoelectric cooling. Recent work (at a wavelength of 1100 nm) has shown direct modulation at 20 Gb/s with an output power of 10 mW (10 dBm) without thermoelectric cooling [3] and is indicative of the trend in laser diode research.

The combination of a particular receiver sensitivity and laser source determines the largest link power budget that can be used. For example, a 1 mW (0 dBm) laser source and a −30 dBm receiver sensitivity can support a link power budget of −30 dB. In Figure 10, this would mean a 16 × 16 switch can be built. On the other hand, if a 10 mW (10 dBm) laser source is used with the same receiver, a link power budget of −40 dB can be supported. This would increase the switch size to 32 × 32. It is interesting to see how the switch size scales as the individual component values improve over time. Figure 11 is a plot of the link power budget as a function of switch size for three different years. The values are taken from Tables 1 and 2. In considering years 2000 and 2002, we only considered improvement in the waveguide attenuation loss and the switch element insertion loss. We assumed the total waveguide length did not change, although it is likely to decrease as the switch elements improve. In addition to the link power budget, two horizontal lines are drawn. These indicate the maximum link power budget that can be supported for a given laser source and receiver sensitivity. For the top line, a 10 mW (10 dBm) laser is coupled with −30 dBm receiver indicating a total link power budget of −40 dB is allowed. This translates into switches of 32 × 32, 128 × 128, and nearly 256 × 256 in 1998, 2000, and 2002 respectively. By increasing the laser source to 100 mW (20 dBm, lower horizontal line in Figure 11), we can build switches of 64 × 64, 128 × 128, and 256 × 256 in 1998, 2000, and 2002 respectively.

6 Network Interface

Extremely fast optical data paths do not improve overall performance if there is a communications bottleneck at the processing nodes. Unfortunately, the current bus-based I/O bandwidth of even high performance workstations is generally unable to keep up with the data rates associated with our target applications. We address this issue by changing the standard memory system design and providing a direct path from the
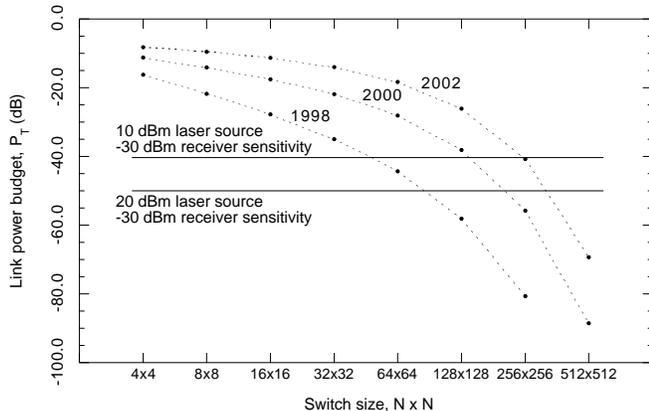
Figure 11: Overall link power budget

Gemini interconnect into the processor's memory, bypassing the I/O bus completely. A block diagram of the *Gemini* network interface is shown in Figure 3.

To provide a fast path for data passing on the optical part of *Gemini*, the traditional main memory-bus path is augmented with dual-port memory modules which can be shared both by the processor bus and the optical interconnect. Muxes 1 through N act to select either the bus or the optical interconnect as the source (or destination) of data to (or from) the memory banks. In addition, a gather/scatter engine is used to collect messages from memory for delivery to the network and distribute messages from the network into memory. Through the use of dedicated gather/scatter engines, network throughput can be maintained even in the case where messages are not tied to sequential blocks of memory, but have some regular block structure. This is common in the applications of interest.

The prototypical use of the gather/scatter engine is to support corner turn operations on STAP applications. Here, the data is commonly block-decomposed matrices, and the communication required from one processor to another consists of some number of fixed size blocks of data that are located a regular distance apart (with a given stride within each block). On conventional message-passing systems, multiple individual messages are required to implement the above transfer. The gather/scatter engine is a finite-state machine that supports this transfer as a single message. In addition to the traditional starting address and message size, the processor provides block size, block separation (space between blocks), and block stride (within a block) information to enable the gather/scatter engine's accessing of memory. An example gather operation is illustrated in Figure 12. The shaded boxes indicate the message to be delivered to the network.

The gather/scatter engine collects the message from memory and delivers it to the network as a contiguous stream of data. The objective here is similar to that found in earlier efforts using permutation networks with SIMD and subsequent parallel processors [2].
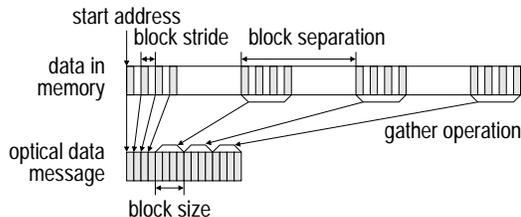

Figure 12: Network interface gather operation

The network interface architecture described above is scalable up to very high bandwidths. For example, consider a 64-bit wide bus, a memory interleaving factor $N = 2$, and a 30 ns (burst mode) memory cycle time. The effective bus throughput is about 2 Gb/s (64 bits/30 ns). In this case, the serial to parallel converter takes data from the optical link and groups it into 128-bit blocks. Memory banks 1 and 2 can then be loaded simultaneously, giving an interconnect bandwidth of 4 Gb/s. With the proper interleaving control, this information can then be accessed by the processor (in 64-bit words) over the bus and the data rate to/from the interconnect is twice as fast as the electrical bus. For higher data rates, this ratio of optical interconnect to bus data rates can be increased by increasing the memory interleaving factor N.

The remaining factor that limits the bandwidth in the optical path is the speed of the laser/detector pair. Commercially available lasers operate at about 2.5 Gb/s, with 10 to 20 Gb/s versions expected to be affordable in the near future. Through the use of wavelength division multiplexing (WDM), it is possible to use multiple lasers simultaneously (each at a different wavelength), thereby providing a completely bandwidth scalable system.

Note that the above design precludes simultaneous communication and computation (since the muxes must remain connected to the optical link during communication due to the throughput requirements of the interconnect). This limitation can be overcome by constructing a pair of communications memories (each appropriately internally interleaved) which are used in a double buffered style. Thus, while communication is occurring to/from one memory, computation can take place out of the other memory. For even higher performance this can be extended to three parallel memory groups, simultaneously supporting inbound communications, computation, and outbound communications.

7 Summary and Conclusions

This paper presented the design of an optically interconnected multiprocessor oriented to applications where performance has been bandwidth limited in conventional multiprocessors. Real-time, space-time adaptive processing (STAP) applications exemplify applications which require the extensive parallel processing of large data volumes derived from sensor arrays. The system is designed to take advantage of the high bandwidth potential of optical interconnects along with the logic, memory, and inexpensive VLSI capabilities of current digital systems.

The system utilizes board-level polymer waveguides to reduce manufacturing costs and commercially available 2×2 LiNbO$_3$ switches. The processor interconnection network has a Banyan topology and is composed of dual optical and electronic networks. The optical data paths (circuit switched) are used for passing large data blocks and the matched electrical data paths (packet switched) are used for setup, tear down, and general control of the optical interconnect and also for short data messages. A optical power budget analysis was presented and, with current optical technologies, up to 32×32 networks can be implemented. Technology projections indicate that within a few years 256×256 optical networks could be built on a single 20 cm by 20 cm printed wiring board. Such a network would have a bandwidth of greater than 20 Gb/s on each link.

In order to match the bandwidth of the optical components with the processor and memory bandwidths available, a sophisticated network interface must be designed. The design presented utilizes dual-port memories and a gather/scatter network. The combination permits reasonable utilization of the optical network and is well suited to the targeted applications. We are currently implementing a scaled down prototype of the system discussed in this paper. The system will contain a 4×4 Banyan switch, off-the-shelf microprocessors, and a limited gather/scatter engine. In experimental work so far, we have successfully communicated (using ATM NICs in PCs as the data sources and sinks) through 4 stages of optical switching.

Acknowledgements

References

[1] T. S. Barry, D. L. Rode, and R. R. Krchnavek. Highly efficient coupling between single-mode fiber and polymer optical waveguides. *IEEE Trans. on Components, Packaging and Manufacturing Technology-Part B: Advanced Packaging*, 20(3):225–228, August 1997.

[2] Kenneth E. Batcher. The Flip network in STARAN. In *Proc. of 1976 Int'l Conf. on Parallel Processing*, pages 65–71, 1976.

[3] K. Czotscher et al. Uncooled high-temperature (130 °C) operation of InGaAs-GaAs multiple quantum-well lasers at 20 Gb/s. *IEEE Photonics Technology Letters*, 9(5):575–577, May 1997.

[4] J. Duato et al. A high performance router architecture for interconnection networks. In *Proc. of the 1996 Int'l Conf. on Parallel Processing*, pages 61–68, 1996.

[5] L. Eldada et al. Low-loss high-thermal-stability polymer interconnects for low-cost high-performance massively parallel processing. In *Proc. of the 3rd Int'l Conf. on Massively Parallel Processing Using Optical Interconnections*, pages 192–205, October 1996.

[6] L. D. Garret et al. Performance of 8-channel OEIC receiver array in 8 x 2.5 Gb/s WDM transmission experiment. *IEEE Photonics Technology Letters*, 9(2):235–237, February 1997.

[7] M. Kajita et al. 1-Gb/s modulation characteristics of a vertical-cavity surface-emitting laser array module. *IEEE Photonics Technology Letters*, 9(2):146–148, February 1997.

[8] Gerd Keiser. *Optical Fiber Communications*. McGraw-Hill, 1991.

[9] Lucent Technologies. Guided wave optical switch products. Preliminary data sheet, 1997.

[10] Craig Lund. Optics inside future computers. In *Proc. of 4th Int'l Conf. on Massively Parallel Processing Using Optical Interconnections*, pages 156–159, 1997.

[11] E. J. Murphy et al. Enhanced performance switch arrays for optical switching networks. In *Proc. of ECIO*, 1997.

[12] J.W. Perry et al. Organic salts with large electro-optic coefficients. In *Proceedings of the SPIE*, pages 302–309, 1991.

[13] Y. Shi et al. Fabrication and characterization of high-speed polyurethane-disperse red 19 integrated electrooptic modulators for analog system applications. *IEEE Journal on Selected Topics in Quantum Electronics*, 2(2):289–299, June 1996.

[14] A. Sneh, J. E. Zucker, and B. I. Miller. Compact, low-crosstalk, and low-propagation-loss quantum-well y-branch switches. *IEEE Photonics Technology Letters*, 8(12):1644–1646, December 1996.

[15] S. Yu et al. A monolithically integrated 1 x 4 switchable photodiode array with preamplifier for programmable frequency filters and optical interconnects. *IEEE Photonics Technology Letters*, 9(5):675–677, May 1997.