The *Gemini* Interconnect:
Data Path Measurements
and Performance Analysis

Ch'ng Shi Baw
Roger D. Chamberlain
Mark A. Franklin
Michael G. Wrighton

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

The *Gemini* Interconnect: Data Path Measurements and Performance Analysis

Ch'ng Shi Baw, Roger D. Chamberlain, Mark A. Franklin, and Michael G. Wrighton
{baw,roger,jbf,wrighton}@ccrc.wustl.edu
Computer and Communications Research Center
Washington University, St. Louis, Missouri

Abstract

The Gemini *interconnect is a dual technology (optical and electrical) interconnection network designed for use in tightly-coupled multicomputer systems. It consists of a circuit-switched optical data path in parallel with a packet-switched electrical control/data path. Here, we present quantitative measurements of optical data path operation from the physical implementation, as well as a discrete-event simulation model of the entire interconnect that includes visualization capabilities. Performance analyses of several aspects of system operation are developed from the simulation model.*

1   Introduction

The *Gemini* interconnect is an experimental implementation of a novel processor-to-processor interconnection network for tightly-coupled multicomputers [1]. It includes an end-to-end optical data path (including switching of the optical signals) for high-bandwidth, large data volume message delivery. The optical switching is accomplished using $LiNbO_3$ electrooptical 2×2 switches [7, 9]. In addition, *Gemini* includes an electrical path (in parallel with the optical path) that both controls the optical path (i.e., setup of the electrooptical switches) and delivers low-latency, small data volume messages.

The *Gemini* interconnect uses a Banyan topology. Although this is a blocking network, it provides the minimum number of switching stages through the network, and has the additional advantage that each signal goes through the same number of switches. An 8×8 *Gemini* network is illustrated in Figure 1.

Due to the absence of buffering in the optical domain, the optical network is circuit switched. This implies that it will perform well for large data volume messages, for which the latency associated with circuit setup and teardown can be amortized over a large message insertion time. By contrast, the electrical network is packet switched. Here, the design can be optimized
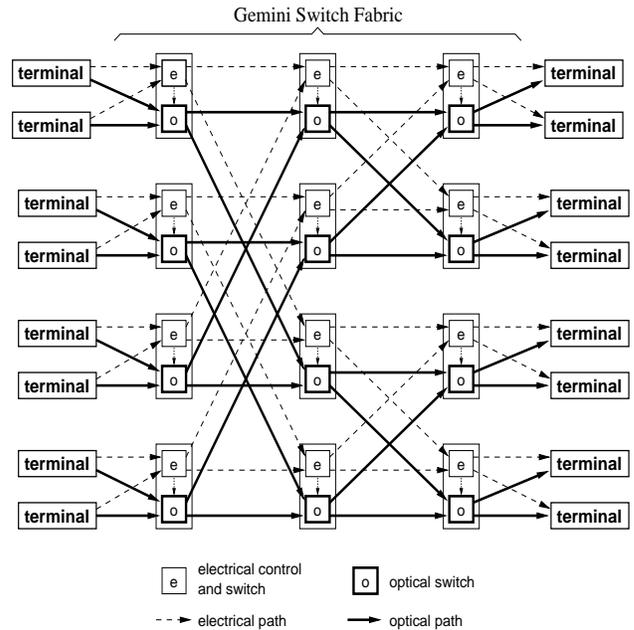

Figure 1: An 8×8 *Gemini* network.

for low-latency delivery of small messages (either data messages or control messages) that do not have significant bandwidth requirements.

The original design concept for *Gemini* was described in [1]. Here, we present measurements from the physical data path, a simulation-based performance analysis tool, data transmission protocol designs, and performance predictions for the interconnect for a number of protocols.

2   Data Path Measurements

Electrooptical 2×2 switching elements are the key devices in the fabrication of the *Gemini* $N \times N$ optical data path. [7, 9]. These switching elements rely on the electrooptic effect (i.e., the application of an electric field to an electrooptical material changes the refrac-

tive index of the material). The result is a 2×2 optical switching element whose state is determined by an electrical control signal. This is illustrated in Figure 2, which shows a switching element in the pass through state as well as in the crossover state.
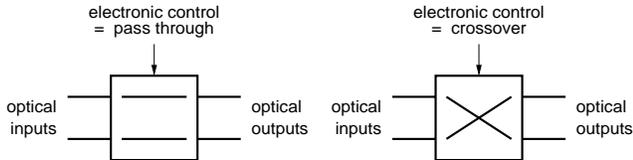

Figure 2: Electrooptical switching elements.

As fabricated by Lucent, these devices are constructed using 1×2 Y-switch elements for crosstalk minimization. They are optimized for use at a wavelength of 1550 nm; however, due to the availability of laser sources, we were restricted to a wavelength of 1300 nm. Here we empirically report on the operation of the LiNbO$_3$ guided wave electrooptical switches operating at 1300 nm.

Table 1 shows the insertion loss for each of 6 switches, measured using a Fotec M712A optical power meter at 1300 nm, and the corresponding data measured by the manufacturer at 1550 nm. The bias control voltage for all of these measurements is +45 V. As can be seen in the data, the performance degrades slightly when operating at 1300 nm, but not significantly.

Table 1: Insertion loss.

switch	measurements at 1300 nm (dB)	manufacturer's data at 1550 nm (dB)
1	-6.6	-5.2
2	-6.6	-4.9
3	-5.1	-5.2
4	-4.7	-5.2
5	-5.5	-5.3
6	-5.6	-5.1

Using 155 Mb/s ATM network interfaces (NIs), the above described electrooptical switches, and fiber links between switching elements, we have demonstrated signal delivery through 4 stages of optical switching (the maximum the power budget will allow for the available lasers and receivers).

Using 622 Mb/s ATM NIs as the signal source, we have demonstrated signal delivery through 3 stages of optical switching (again, the maximum that the power budget allows). Figures 3 and 4 show the signal both immediately out of the source (Figure 3) and after one stage of optical switching (Figure 4). Note that although the height of the eye pattern has been diminished (as predicted by the insertion loss of the optical switch, note the change in vertical scale), the width of the eye is essentially the same (to the resolution of the measurement instrumentation), indicating that the optical switching elements do not significantly degrade the signal shape. This is an important consideration for scaling up to higher bit rates.

Finally, at 2.4 Gb/s, we have successfully demonstrated signal delivery through three stages of optical switching. Here, the signal source was a SONET line card for the Washington University Gigabit Switch [2]. All of the above experiments were performed at a wavelength of 1300 nm.
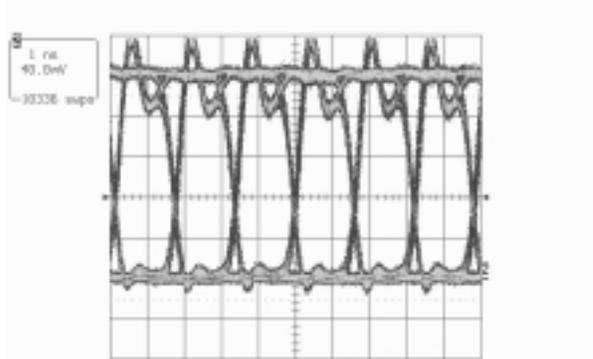
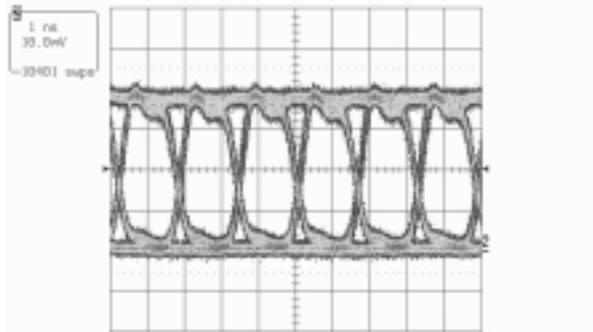

Figure 3: Eye pattern into optical switch.

Figure 4: Eye pattern out of optical switch.

An important question of interest to system designers is the degree to which the details of electrooptical switch operation can be encapsulated into a small set of operating specifications. Specifically, we present empirical data that addresses the question of whether or not the switch can effectively be modeled (for system design purposes) as a linear system with a given in-

sertion loss (i.e., non-linear effects can be ignored and losses are simply additive).

We are interested in whether non-linear effects in the switch cause sufficient signal degradation that the bit-error rate at the receiver is larger than would be predicted with a linear model. Assuming Gaussian noise, the bit-error probability, $P[\mathcal{E}]$, for an optimum receiver is given as [11]:

$$P[\mathcal{E}] = Q\left(\sqrt{E_s/\sigma^2}\right) \qquad (1)$$

where

$$Q(\alpha) = \frac{1}{2}\left[1 - \mathrm{erf}\left(\frac{\alpha}{\sqrt{2}}\right)\right] \qquad (2)$$

E_s is the energy in the signal, and σ^2 is the variance in the noise.

Figures 5 and 6 show both a waveform and noise histogram for a high-level signal and a low-level signal, respectively, directly from the laser source. The histograms are over the portion of the waveform shaded at the top of each figure (the region 1.4 ns to the right of center, between the center line and the curser line). The detector sensitivity (into the oscilloscope) is 1 V/mW, and the sources are operating at 155 Mb/s.

Figure 7 shows a waveform and noise histogram for a high-level signal at the output of the optical switch. Both the mean signal level and the standard deviation of the noise have been attenuated in a linear fashion, indicating that non-linear effects are minimal. Although not shown here, a similar effect occurs for the low-level signals.

The above measurements were taken with the unused input to the switch dark. When another signal source (similar to Figures 5 and 6 but at an arbitrary phase relationship) is passed through the switch via the unused input/output pair, the result (for the output of interest) is shown in Figure 8. As can be seen, both the waveform shape and measured statistics do not differ in any significant way from the previous figure. This indicates that (for input signals at similar power levels) crosstalk is not a significant problem for these switches. Combine this with the fact that all paths in the network traverse the same number of switches, and we conclude that crosstalk within the switches is unlikely to be an issue for the *Gemini* system.

3 Performance Analysis Tools

In the above section, we investigated the physical optical paths, delivering data from point A to point B. Starting with this section, we widen our interest to address system-level performance issues as impacted by both raw data rate and data transmission protocol
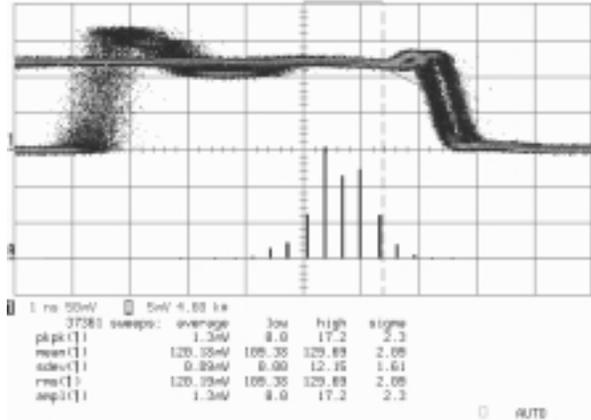

Figure 5: Waveform and histogram of signal levels for high bit from laser source.
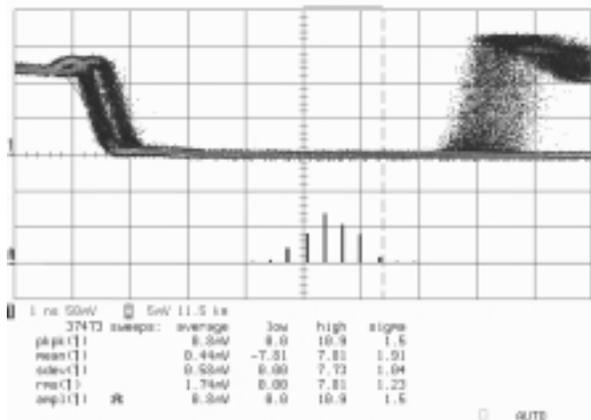

Figure 6: Waveform and histogram of signal levels for low bit from laser source.

choices. The current section describes the performance analysis tools employed, and the sections that follow present the performance results.

The basis for the performance results presented in this paper is a discrete-event simulation model of the *Gemini* interconnect and an associated visualization tool that allows one to both verify correct operation of the simulation as well as help understand overall system behavior. The discrete-event simulations are implemented using the MODSIM III language. The visualization tool is driven using trace data from the simulation and is implemented in Java (primarily for portability reasons).

The terminals connected to the *Gemini* network are modeled as general purpose processors with electrical and optical interfaces. Figure 9 depicts the model of a terminal. The data message generator can be used to generate synthetic load for simulation purposes. Ap-
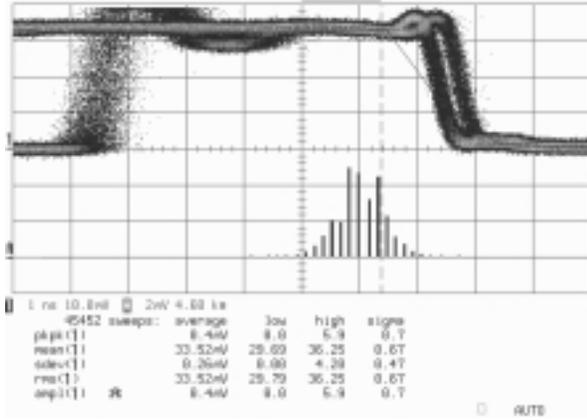
Figure 7: Waveform and histogram of signal levels output from switch for high bit without crosstalk.
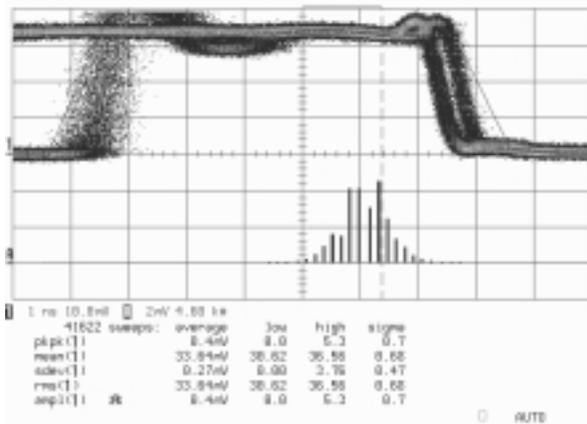

Figure 8: Waveform and histogram of signal levels output from switch for high bit with crosstalk present.

Figure 9: Model for a terminal attached to the network.
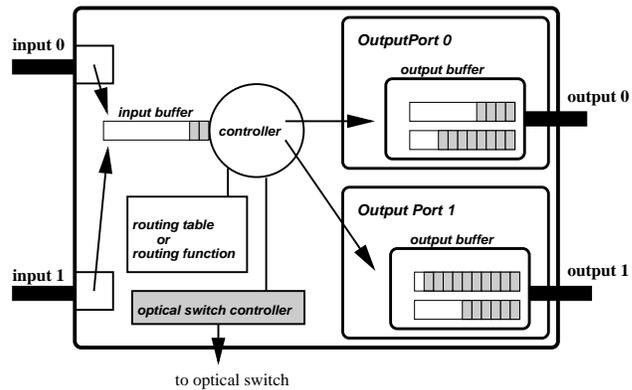
Figure 10: Model for the *Gemini* electrical switch.

plications can be modeled into the CPU module. Since the CPU module is the only module that consumes received data messages, all data messages are routed to the CPU module. Network control signals are assumed to be processed at line rate. Hence there is no input buffer for the terminal. The terminal has separate output buffers for messages intended for different networks. The controller marked 'A' dispatches incoming packets according to packet type. The controller marked 'B' dispatches outgoing traffic according to message type and length.

Figure 10 shows the model of a *Gemini* 2×2 electrical switch. The electrical switch has a shared input buffer and separate output buffer at each output. A routing function module informs the controller where to forward a packet as well as how to control its companion optical switch when a path setup request is being processed.
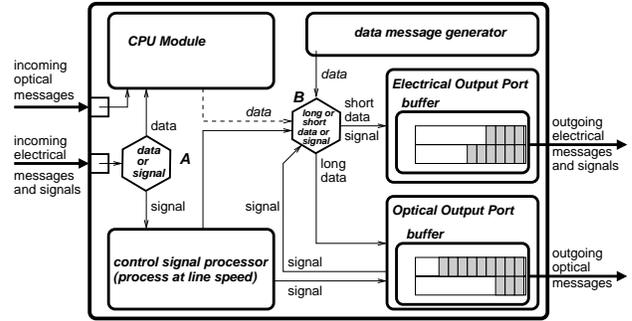
In the simulation, the optical switch and electrical switch are modeled as one object. The links that connect the switches are modeled as one link entity with two channels, one channel carries the electrically switched traffic, the other the optically switched traffic. Further details on the implementation issues and object structure of the discrete-event simulation models are described in [5].

The original protocol for *Gemini* proposed in [1] required a negative acknowledgement signal to be back-propagated should an attempt to setup an optical path fail. As such, electrical switches must be capable of bidirectional communications. This protocol required a wait time proportional to maximum electrical network packet size and network size between sending a path setup request and transmitting optical data. It does not allow different optical path setups to be attempted in parallel. In the next section, we propose a modified protocol that does not have these restrictions and requires only unidirectional communications within the network. Since none of the protocols proposed in this paper require bidirectional communication, we model the electrical switches as only capable of unidirectional communication.
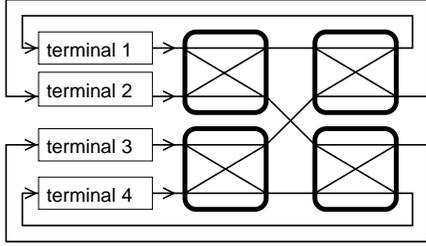
Figure 11: Each output is routed back to a corresponding input terminal in a 4×4 network.

Although it is customary to draw input terminals (senders) on one side and output terminals (receivers) on the other, with a unidirectional communication network, it is assumed that each output is routed back to the input terminal as shown in Figure 11.

To support both the verification of the simulation models as well as improved understanding of the operation of the system as a whole, a visualization tool is used. The visualization tool is driven from a static topology description file and trace data derived from the simulation execution. The topology description defines the structure of the network: terminal objects, switching node objects, queues within terminals and switching nodes, and links between objects. This structure closely follows the form of the diagram of Figure 1. Each of the switching nodes consists of an electrical switch (on top) and an optical switch (below). Links (both electrical and optical) form connections between the terminals and the switching nodes. Within each of the terminals and switching nodes, the message queues are represented graphically. Different message types (e.g., *setup*, *teardown*, *data*) are represented by distinct colors. When in use, links take on the color of the message type in transit.

Simulator derived trace data encapsulates the dynamic activity present in the network. This reflects the state of the queues, links, and optical switches (i.e., either pass through or crossover). A complete design description of the visualization tool is given in [12].

4   Interconnect Performance

This section describes the basic *Gemini* transmission protocol adapted from that originally proposed in [1] and presents performance results for the resulting system. In contrast to the original protocol, our protocol uses positive acknowledgement. For simplicity, we first assume that there is no application data sent via the electrical network – the electrical network is solely used to send network control signals. Transmission of application data over the electrical network will be ad-
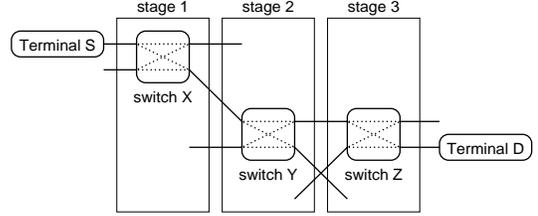

Figure 12: Terminals S and D connected via a 3-stage network.

dressed later.

Suppose a terminal S wants to send a message to another terminal D. An optical path needs to be setup so that the optical signal transmitted by S can be sensed by D's detector. Suppose there are three stages of switches in the network as shown in Figure 12.

To set up an optical path, S sends a *setup* signal to D using the electrical network. The *setup* signal has to pass through switches X, Y, and Z before it reaches D. If all of the X, Y, and Z switches grant the *setup* request, D will send an *ackSetup* signal to S. S can start optical transmission after it receives D's *ackSetup* signal. If one of the X, Y, and Z switches cannot grant S's setup request, the switch will set a *blocked* flag in the *setup* signal. If D receives a *setup* signal with the *blocked* flag set, D will send S a *block* signal to S. S will send a *teardown* signal to D if it receives a *block* signal or after it has finished its optical transmission. The *setup* signal tells the switches to maintain their state and hence an optical path. The *teardown* signal tells the switches that they no longer need to maintain their state. We call this protocol the basic *setup-teardown* protocol.

The protocol requires that a terminal does not send another setup request until it has sent a *teardown* signal following the receipt of a *block* or *ackSetup* signal. Thus, no switch needs to keep track of more than one outstanding setup request per terminal.

We define the following quantities to aid our analysis of the *Gemini* network.

- N: Network size. An $N \times N$ Banyan network has N inputs and N outputs. We consider only network sizes where N is a power of 2.

- BW_e: Electrical link bandwidth.

- BW_o: Optical link bandwidth.

- l_{sig}: Length of a control signal.

- $\overline{l_o}$: Mean length of messages sent on optical network.

- t_{sig}: Time needed to send a control signal on electrical link, defined as $t_{sig} \equiv l_{sig}/BW_e$.

- α_{sig}: Time needed to process a signal by an electrical switch expressed as a factor of t_{sig} (e.g., if t_{sig} is 8 clock ticks and the time needed to process a control signal is 4 clock ticks, then $\alpha_{sig} = 0.5$).

- $Tput_{o-max}$: Maximum optical network throughput.

- η_{o-max}: Optical network maximum utilization efficiency.

The maximum utilization efficiency, η_{o-max}, is the portion of time the optical network spends transmitting data in the absence of contention and blocking under the infinite load assumption (i.e., there are always optical messages to be sent at every sender). As these two conditions are unlikely to both simultaneously occur, this upper bound is not an achievable operating point with realistic traffic patterns.

4.1 Control Signal Latency Analysis

Critical to *Gemini*'s optical network performance are the *setup*, *ackSetup*, *block* and *teardown* signals. Thus we concern ourselves initially with the latency of these signals. The minimum latency for a control signal in the absence of congestion can be obtained as

$$T^{sig}_{min-lat} = t_{sig}((1 + \alpha_{sig}) \log_2 N + 1) \qquad (3)$$

The terminal network interface takes t_{sig} to insert the control signal into the network. In the absence of congestion, the signals propagates through the $\log_2 N$ stages without queueing delay. Each stage takes $(1 + \alpha_{sig}) t_{sig}$ to forward the signal, $\alpha_{sig} t_{sig}$ for message processing and t_{sig} for insertion into the next link. Thus we arrive at Equation 3.

Using the basic setup-teardown protocol described above, optical transmission cannot commence until an *ackSetup* signal is received. Thus it is of interest to analyze how long it would take to receive a corresponding *ackSetup* after a *setup* signal was sent. Without contention, the minimum wait period, RTT^{sig}_{min}, between sending a *setup* and receiving an *ackSetup* is

$$RTT^{sig}_{min} = 2T^{sig}_{min-lat} = 2t_{sig}((1+\alpha_{sig}) \log_2 N+1) \quad (4)$$

4.2 Optical Network Peak Performance Analysis

Using the basic *setup-teardown* protocol, the optical network achieves the highest throughput when no connection setup request is blocked and no *setup*, *ackSetup*, or *teardown* signal experiences any queueing delay.

Specifically, maximum optical network throughput is achieved when every optical path setup process is completed in RTT^{sig}_{min}. Since every path setup process (except the first) has to be preceded by a *teardown* signal, the minimum time between the completion of a message and the beginning of another message sent by the same network interface, $T^o_{min-idle}$, is

$$
\begin{aligned}
T^o_{min-idle} &= RTT^{sig}_{min} + t_{sig} \\
&= (2(1 + \alpha_{sig}) \log_2 N + 3)t_{sig} \quad (5)
\end{aligned}
$$

Given that the average optical message length is $\overline{l_o}$, the average time needed to send one optical message T^o_{busy}, is

$$T^o_{busy} = \overline{l_o}/BW_o \qquad (6)$$

Thus the maximum optical network utilization efficiency, η_{o-max}, is

$$\eta_{o-max} = \frac{T^o_{busy}}{T^o_{busy} + T^o_{min-idle}} =$$

$$\frac{\overline{l_o}/BW_o}{\overline{l_o}/BW_o + (2(1 + \alpha_{sig}) \log_2 N + 3)l_{sig}/BW_e} \qquad (7)$$

The maximum optical network throughput occurs when each of the N inputs is communicating in a non-blocking fashion to each of the outputs.

$$Tput_{o-max} = N \eta_{o-max} BW_o =$$

$$\frac{N \overline{l_o}}{\overline{l_o}/BW_o + (2(1 + \alpha_{sig}) \log_2 N + 3)l_{sig}/BW_e} \qquad (8)$$

We further define the following:

- Γ: Average optical message length to control signal length ratio. Defined as $\Gamma \equiv \overline{l_o}/l_{sig}$.

- Θ: Ratio of optical link bandwidth to electrical link bandwidth: $\Theta \equiv BW_o/BW_e$.

The maximum utilization efficiency expression in Equation 7 can be rewritten using Γ and Θ as follow:

$$\eta_{o-max} = \frac{\Gamma}{\Gamma + \Theta(2(1 + \alpha_{sig}) \log_2 N + 3)} \qquad (9)$$

Optical links are idle in between messages due to setup delay. Thus intuitively one would expect that the longer the optical messages, the less frequently the optical links are idle. Thus longer optical messages (i.e., larger Γ) should improve optical network utilization efficiency. One would also expect that the faster the electrical control network (i.e., smaller Θ), the faster one can setup an optical path, and the more efficiently one can utilize the optical network. Furthermore, the

larger the network, the longer it takes to setup an optical path. Thus one should also expect that all else being equal, the larger network will be less efficient in terms of optical network utilization. Equation 7 supports all of the above.

The four graphs shown in Figure 13 plot η_{o-max} as a function of Γ and Θ for different network sizes N using Equation 9 with $\alpha_{sig} = 1$. These plots confirm the above intuition.
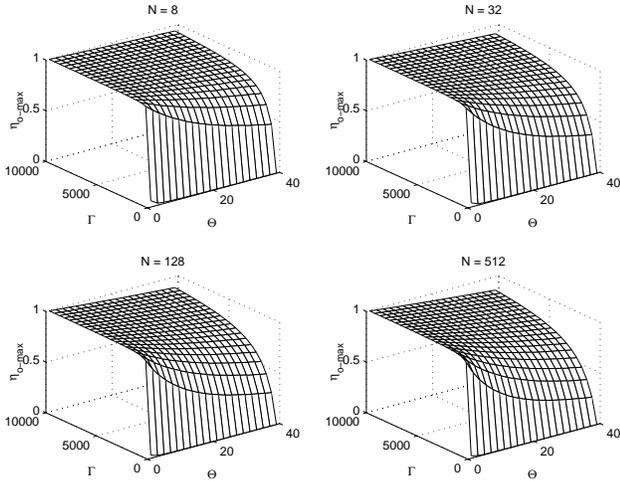
Figure 13: Plots of η_{o-max} as a function of Γ and Θ.

All the above analyses assume that optical switching time is less than $T^{sig}_{min-lat}$.[1]

4.3 Simulation Results

Figures 14 to 16 show the simulation results for four *Gemini* networks using the basic *setup-teardown* protocol. In these simulations, all application data messages are sent via the optical network. The parameters chosen for the simulated networks are such that $\Gamma = 16384$, $\Theta = 12$, and $\alpha_{sig} = 1.25$. According to Equation 9, we should be able, under no-contention conditions, to achieve 99.20%, 98.91%, 98.63%, and 98.34% utilization respectively for the 4×4, 8×8, 16×16, and 32×32 *Gemini* networks.

The four networks were simulated using the same set of parameters. Two sets of simulations were performed on these networks. The first set of simulations uses constant length messages. The other set uses variable length messages where the lengths are exponentially distributed. The mean of the exponential distribution is the same as the constant length used in the first set

---

[1] The LiNbO$_3$ optical switch requires a 45 V swing to change state [7]. A prototype driver circuit (not yet optimized for speed) developed locally performs such a swing in 10 μsec.
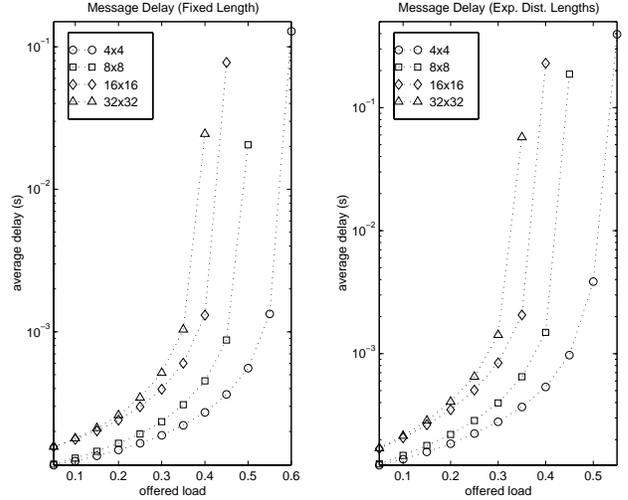
Figure 14: Average optical message delay using the basic *setup-teardown* protocol.

of simulations. All parameters are otherwise identical in the two sets of simulations. Messages are generated according to an independent and identically distributed Poisson process. Their destinations are uniformly distributed to all outputs. In all the figures, the load axes are normalized to the theoretical maximum throughput calculated using Equation 8.

Figure 14 plots the average delays experienced by the optical messages. It shows that the networks approached saturation in the 35% to 60% normalized load region. The largest network saturates with the lowest normalized load. Since utilization cannot exceed load, we conclude that the maximum utilization of the optical network simulated is in the 35% to 60% region. As one would expect, the constant length messages have a lower delay than the messages with exponentially distributed lengths. As will be described below, this poor performance is due to blocking in the optical network.

Plotting the delay experienced by the control signals (see Figure 15) shows no sign of congestion in the electrical network (i.e., control signals are traversing the electrical network in minimum time). Thus we conclude that the contention free assumption for the electrical network has been met.

The analysis to this point has assumed that all application messages are delivered via the optical network. In the *Gemini* conceptual design, it is intended that small application messages (those requiring low latency but not high bandwidth) be delivered via the electrical network. When small application messages use the electrical network, the performance analyses presented here are all still valid, as long as the quantity of application messages does not significantly impact the lack
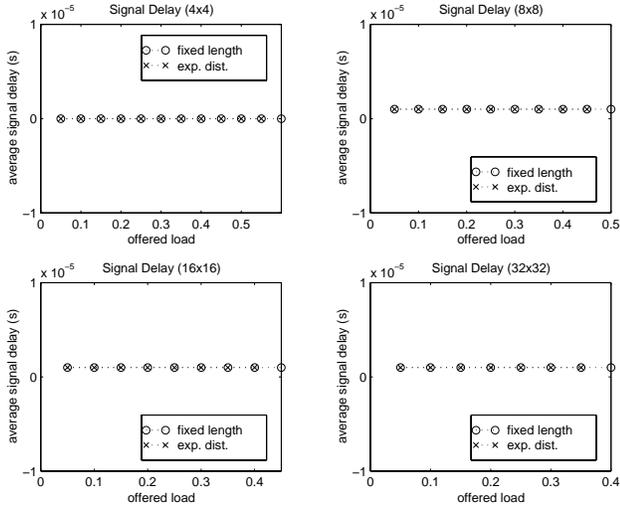
Figure 15: Average control signal delay using the basic *setup-teardown* protocol.

Figure 16: Network blocking rate using the basic *setup-teardown* protocol.

of congestion on the electrical network.

The other major assumption in the development of Equation 7 is that the optical network be free of blocking. By plotting the blocking rates (i.e., how many blocked setup requests are received per unit time) in Figure 16, we see that blocking is the primary cause of the poor throughput achieved by the simulated networks. In addition, the blocking rate increases significantly with network size.

5 Improving Throughput Using Virtual Output Queues

We see from the simulation results presented in the previous section that the optical network becomes saturated when offered load is only in the 35% to 60% range. The primary cause for the poor performance is excessive blocking in the Banyan network using the simple *setup-teardown* protocol.[2]

Consider the 4×4 network depicted in Figure 17 for example. Terminal $S1$ is sending an optical message to terminal $D3$, and terminal $S4$ is sending an optical message to terminal $D2$. This connection requires that switches W and X be in the crossover state and switches Y and Z be in the pass through state to sustain the optical paths in use.

In the queues of terminals $S2$ and $S3$ are messages to various destinations. The head-of-line message in $S2$ cannot be sent to $D4$ due to blocking. Similarly, the

head-of-line message in $S3$ cannot be sent to $D1$. Thus the $S2$-to-$D1$ and $S3$-to-$D2$ optical paths are wasted. These idle paths could have been used if $S2$ is allowed to send the second message in its queue to $D1$, and $S3$ is allowed to send the second message in its queue to $D2$.
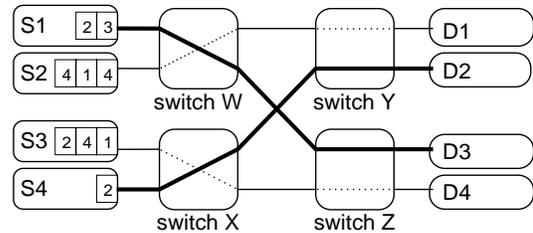
Figure 17: Head-of-line blocking in a 4×4 banyan network.

A way to avoid such inefficiency is to allow a source to explore multiple optical paths in parallel during the setup process. One approach is to have messages stored in multiple queues according to their destinations. Such queues are termed *virtual output queues (VOQ)* [10]. The *Gemini* VOQ protocol is similar to the basic *setup-teardown* protocol except that terminals are allowed to send a *setup* signal for every non-empty VOQ. The initiation of a path setup for each non-empty queue is called a *setupBurst*.

A set of simulations using the same parameters described in Section 4.3 was run with the VOQ protocol. Figure 18 plots the average delays experienced by the optical messages. We see that the optical network can provide close to 100% throughput using the VOQ

---

[2]Even for non-blocking switching networks (e.g., crossbar), with simple FIFO queueing at the input, throughput is still limited to 58% of full capacity of the switching network assuming random, homogeneous traffic [6].
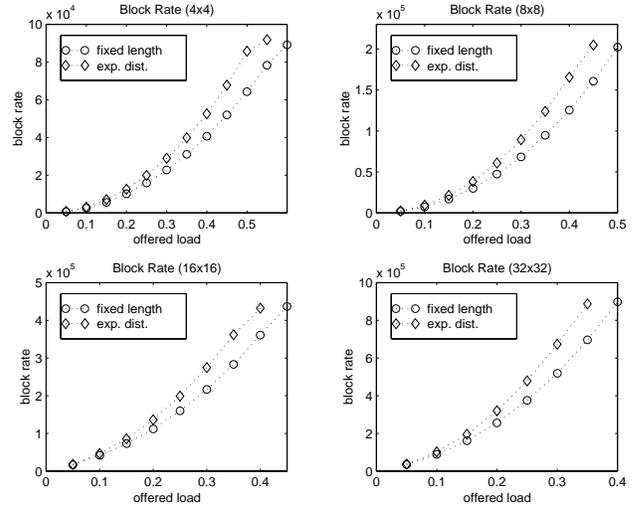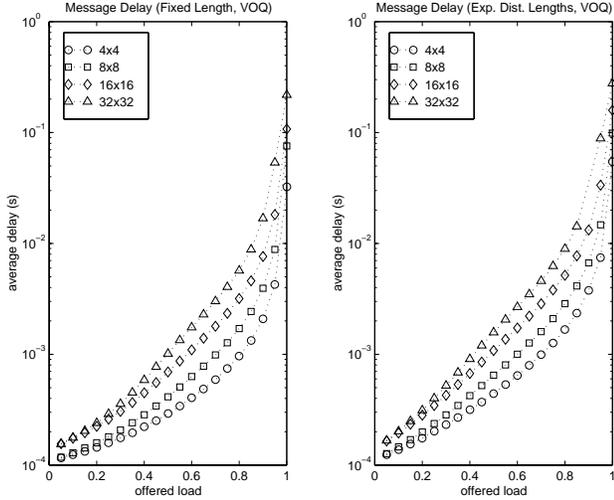
Figure 18: Average optical message delay using the VOQ protocol.

protocol.[3] Figure 19 plots the average delays experienced by the control signals on the electrical network. We see that for the parameters chosen, sending multiple setup requests does not lead to congestion in the electrical network.
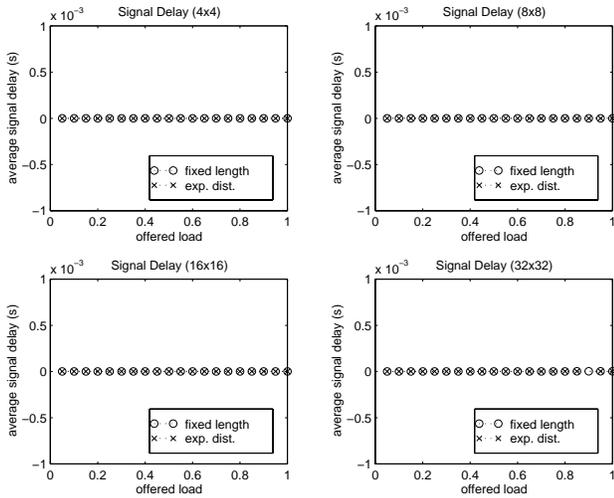

Figure 19: Average control signal delay using the VOQ protocol.

Figure 20 plots the number of blocked setup requests received per second over the entire network. In com-

[3]McKeown et al. have proven in [8] that 100% throughput is achievable in a non-blocking, input-queued switch using a non-FIFO queueing scheme such as VOQ assuming random, homogeneous traffic. It remains to be seen whether such performance is achievable in a blocking network such as the Banyan network used in *Gemini*.

parison, blocking rates in VOQ networks are higher than non-VOQ networks. This is to be expected since the VOQ protocol sends multiple requests simultaneously, knowing that some will be blocked. Figure 20 shows that the blocking rate *decreases* when the load is high. While this may seem counter intuitive, it can be explained as follows.
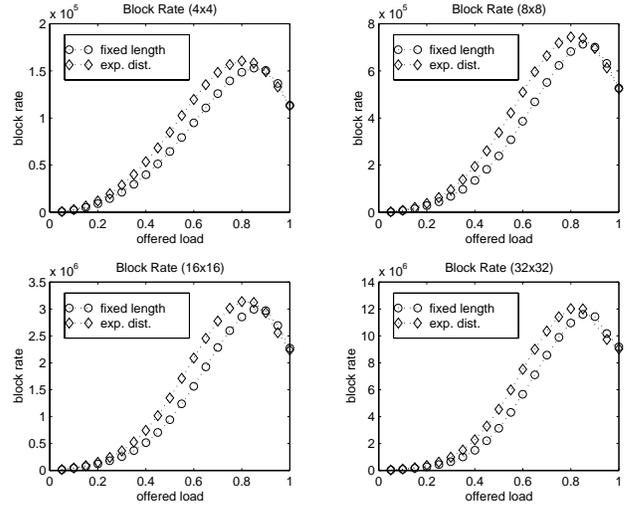

Figure 20: Network blocking rate using the VOQ protocol.

With low load, network contention is low. Thus the blocking rate is low. As the offered load increases, contention increases as well. Thus leading to a higher blocking rate. Note that with medium load, not all queues in a sender are full most of the time. It is still likely that when a *setupBurst* attempts to setup optical paths for all the non-empty queues, all setup requests will fail and thus trigger a *setupBurst* retry after a brief pause. However, if the offered load reaches a certain level, then it is more likely that all queues are full most of the time. When all queues are full, the VOQ *setupBurst* explores all possible paths in the network, thus virtually guaranteeing at least one successful path setup per *setupBurst*. A successful setup prevents further retries for one message time, leading to a reduced blocking rate.

6 Conclusions

The work described in this paper allows us to come to a number of conclusions. The first set of conclusions concerns the physical implementation:

- The *Gemini* network is a feasible design.
- The LiNbO$_3$ electrooptical switches perform well. Their performance at 1300 nm is only slightly worse than at their design point of 1550 nm.

- For system-design purposes, it is reasonable to model the electrooptical switches as linear devices. The theoretical bit-error rate follows a linear model of the switch.
- Crosstalk is not a serious issue, as long as the topology maintains common path lengths from source to destination.

The second set of conclusions come out of the performance analysis:

- For all of the protocols investigated, contention in the electrical network is not significantly present. Control messages traverse the network with minimum latency.
- The straightforward setup-teardown protocol does not perform well due to blocking in the optical network.
- Virtual output queueing goes a long way toward addressing the throughput concerns.

There are a number of issues that warrant further investigation:

1. The current implementation uses fiber between electrooptical switches. We eventually plan to use polymer waveguides to provide the switch-to-switch optical links.
2. The virtual output queueing setup protocol described here is inherently an unfair protocol, since there is hysteresis present in the optical switch configurations. Algorithms to address this problem are described in [3] and [4].
3. There are a number of issues associated with efficiently implementing the virtual output queueing protocol (e.g., memory management in the queue structures). These are investigated in [3].
4. Using the simulator, it is possible to quantify the effects of delivering small application messages via the electrical network. This performance study still needs to be accomplished.

Finally, the message traffic models used here are fairly simple. We are currently in the process of developing application models that provide more realistic message traffic information. They also incorporate the computational requirements of the application, which can be simulated using the existing terminal model (CPU Module). The results of this analysis will give performance figures tied to real application requirements.

Acknowledgements

References

[1] R. Chamberlain, M. Franklin, R. Krchnavek, and B. Baysal. Design of an optically-interconnected multicomputer. In *Proc. of 5th Int'l Conf. on Massively Parallel Processing Using Optical Interconnections*, pages 114–122, June 1998.

[2] T. Chaney, J.A. Fingerhut, M. Flucke, and J.S. Turner. Design of a gigabit ATM switch. Technical Report WUCS-96-07, Dept. of Computer Science, Washington University, St. Louis, MO, 1996.

[3] Ch'ng Shi Baw. Design, analysis, and simulation study of optical interconnection networks. Master's thesis, Washington University in Saint Louis, 1999.

[4] Ch'ng Shi Baw, Roger D. Chamberlain, and Mark A. Franklin. Fair scheduling in an optical interconnection network. In *Proc. of MASCOTS*, October 1999.

[5] Ch'ng Shi Baw and M.A. Franklin. An interconnection network simulator. Technical Report WUCCRC-99-03, Computer and Communications Research Center, Washington University, St. Louis, MO, 1999.

[6] M. Karol, M. Hluchyj, and S. Morgan. Input versus output queueing on a space division switch. *IEEE Transactions on Communications*, 35(12):1347–1356, 1987.

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

[8] N.W. McKeown, V. Anantharam, and J. Walrand. Achieving 100% throughput in an input-queued switch. In *Proc. of IEEE Infocom*, March 1996.

[9] E.J. Murphy, T.O. Murphy, R.W. Irvin, R. Grencavich, G.W. Davis, and G.W. Richards. Enhanced performance switch arrays for optical switching networks. In *Proc. of ECIO*, April 1997.

[10] Balaji Prabhakar and Nick McKeown. On the speedup required for combined input and output queued switching. Technical Report CSL-TR-97-738, Stanford University, November 1997.

[11] J.M. Wozencraft and I.M. Jacobs. *Principles of Communication Engineering*. John Wiley & Sons, 1965.

[12] Michael G. Wrighton. Visualization tool for optical networks. Technical Report WUCCRC-99-02, Computer and Communications Research Center, Washington University, St. Louis, MO, 1999.