VLSI Photonic Ring Interconnect for Embedded Multicomputers: Architecture and Performance

Roger Chamberlain
Mark Franklin
Abhijit Mahajan

Roger Chamberlain, Mark Franklin, and Abhijit Mahajan, “VLSI Photonic Interconnect for Embedded Multicomputers: Architecture and Performance,” in *Proc. of the ISCA 14th Int'l Conf. on Parallel and Distributed Computing Systems*, August 2001, pp. 351-358.

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

VLSI Photonic Ring Interconnect for Embedded Multicomputers: Architecture and Performance

Roger Chamberlain, Mark Franklin, and Abhijit Mahajan
Computer and Communications Research Center
Washington University, St. Louis, Missouri

Abstract

This paper presents both an architectural design and performance analysis of a multicomputer interconnection network based upon the use of optical technology. The system uses Vertical Cavity Surface Emitting Lasers (VCSELs) and free space optics to implement a physical ring topology logically configured as a multiring. A novel network and memory interface architecture is proposed to enable the processing nodes to exploit the bandwidth made available by the optical interconnect. The Deficit Round Robin (DRR) protocol is adapted to the ring for media access. A discrete-event simulation model of the interconnect and protocol is used to examine performance issues such as throughput, latency and fairness.

1 Introduction

Recent advances in VLSI photonic technology permit the design and implementation optical interconnection networks with terabit per second bandwidth capacity. A physical ring architecture (organized as a multiring) that exploits the high bandwidth provided by this new technology is described here and its performance quantified. The design is targeted at multicomputer systems that require frequent, massive data transfer among processors, from input sensors, and to output devices. Examples of embedded systems with these requirements include space-time adaptive processing [8] and real-time image processing applications. Use of the interconnect in the area of high performance router applications is also anticipated.

At the heart of the interconnection network is a VLSI photonic device based on the use of an $M \times M$ array of Vertical Cavity Surface Emitting Laser (VCSEL) and detector pairs. Each VCSEL-detector pair is capable of operating at rates exceeding 1 Gb/s. Prototype interconnects with $M = 16$ have been constructed [12], and designs with $M = 32$ will soon be available. This will provide 1024 VCSEL-detector pairs per chip and provide a raw bandwidth in excess of 1 Tb/s.

The architecture described here uses 2-D arrays of VCSELs, devices which are available commercially. Additionally, the optical components may be bump bonded directly to a CMOS chip. The latter may be an ASIC containing a full processor or portions of an intelligent router. Optical connections between ICs can be implemented with flexible light pipes [6]. The potential here for implementing high performance interconnection networks both for parallel processor implementations and for use in data communication routers is enormous. However, the properties of the optical devices and their high data rates have architecture implications, and some of the lower-level design constraints force one to deal with certain issues, such as fairness, within the higher-level protocol structure.

This paper presents an architecture well suited to VCSEL array technology, examines certain basic performance metrics and considers a protocol design which enforces fair access to the interconnection fabric. Section 2 describes example applications that could benefit from such a system. Section 3 gives an introduction to the optical technology used in the interconnect. Section 4 presents the architecture of the system. Section 5 describes performance results from a discrete-event simulation model of the system, and Section 6 addresses an issue that is crucial to the usefulness of the system: fairness. Section 7 summarizes our results and concludes the paper.

2 Applications

The system described in this paper represents a significant gain in interconnection network bandwidth provided between the processors of a multicomputer. As such, the type of application that could benefit from its existence is one in which communications performance is paramount (i.e, it is communications bandwidth limited).

[1] This research is supported in part by the DARPA VLSI Photonics Program under grant DAAL01-98-C-0074. Abhijit Mahajan is currently with Rambus Inc.

There are a significant number of embedded systems applications for which this limitation is present. Large volumes of sensor-derived data are often processed into a smaller data set and presented to a user or automated control system. Critical communications performance bottlenecks typically occur during data transpose, or "corner turn" operations, and memory access patterns during such operations are often not sequential [8].

Two applications that fit this model are the Automatic Target Recognition (ATR) problem using Synthetic Aperture Radar (SAR) images [11] and the image recontruction problem associated with medical ultrasonic imagery [5]. It is desired that both of these applications execute in real time. However, the bandwidth capabilities of currently available multicomputer systems will not support the necessary data throughput.

The problem of automatically identifying an unknown target from a sub-region (chip) within a SAR image is both computationally intensive and has extremely high data requirements. In most formulations, a chip is compared against the entries in a target database that spans a range of parameters, including target class, configuration, location, orientation, articulation, and obscuration. Estimates are that the search space may need to consist of over 900 million discrete combinations of these parameters in order to deliver useful results. In an embedded parallel system, it is not practical to replicate this database at each computing node, necessitating high-bandwidth access of the database across the interconnection network.

In [11], the size of the database, throughput of the system, and quality of the results are related to one another. An example system configuration containing 64 processors can achieve better than 95% recognition accuracy at a chip rate of 10,000 chips/second; however, this requires a sustained data throughput of 3.2 Tb/s.

Medical ultrasonic imaging is widely used because of its low cost, use of non-ionizing radiation, and ability to operate in real time. Linear transducer arrays enable the user to focus-and-steer the ultrasonic beam over the image region. Real-time operation, however, limits the spatial range over which the beam can be in focus. In contrast, synthetic focus imaging separates data aquisition from image formulation so that a real-time image can be in focus at every pixel.

Parallel implementations of the image reconstruction computation are investigated in [5] and [9], with real-time performance for a 32 element transducer array achievable using 100 processors (assuming 4-way superscalar processors and prefetch to L2 cache). The data bandwidth coming off the sensor array, however, is 400 Gb/s.

3 Optical Technology

The enabling technology for this system is the availability of 2-dimensional arrays of VCSELs and detectors bonded to silicon circuitry [7]. The union of silicon processing with GaAs-based optoelectronics provides a powerful combination, significantly increasing the communications bandwidth available off-chip.

Prototype interconnects have been constructed with 16×16 arrays of VCSELs and photodetectors [12]. In this system, the VCSELs arrays and photodiode arrays were flip-chip bonded to a CMOS chip using heterogeneous integration techniques. Although the demonstration of [12] used bulk optics to deliver light between ICs, the free space optical path for a viable system design could use either a rigid optical link [1] optimized to be tolerant to misalignment (useful for chip-to-chip links on a board), or a flexible fiber imaging guide [6] (useful for board-to-board links).

The large number of VCSEL-detector pairs in the optical interconnect suggests the use of space-division multiplexing to separate the individual channels in the multiring topology. Figure 1 illustrates the allocation of VCSELs and detectors for a four channel system with 16×16 arrays of optical elements. Here, one quarter of the elements are used for each channel. If the individual element data rates are 1 Gb/s, this yields $16^2/4 = 64$ Gb/s for each channel.

4 System Architecture

The overall system considered here is an embedded multicomputer. It is intended that final design decisions (dimensioning, configuring, etc.) be guided by a specific application or set of applications, rather than attempting to be completely general purpose. Each node of the multicomputer has a number of processors, P, local memory, and is arranged in a symmetric multiprocessing configuration. Some fraction of the local memory is designated as *communication memory*, as described below in Section 4.2, and is the primary data interface to the optical interconnection network.

The entire multicomputer has N compute nodes, providing a total of PN processors. Communication between nodes is provided by the optical interconnect. In addition, input devices (e.g., sensors) and/or output devices (e.g., displays) might also be present as nodes on the interconnect.

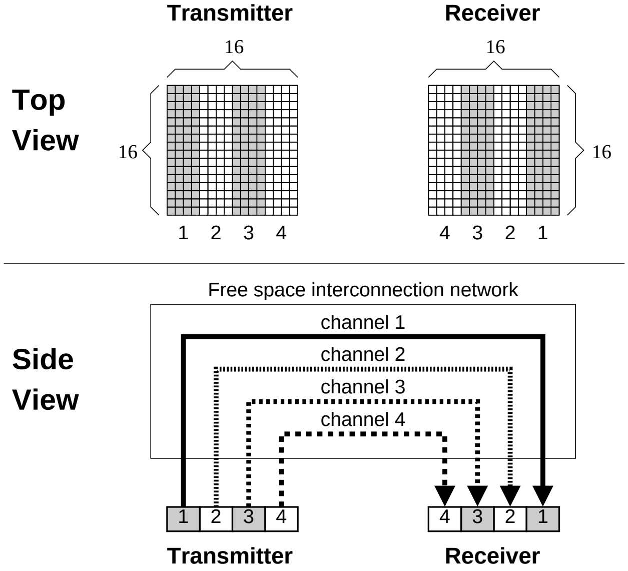

Figure 1: Allocation of VCSEL-detector pairs to a four channel system. 16×16 VCSEL-detector arrays are used, with a 4×16 array allocated to each channel.

The free space optical technologies described here are most cost-effective when used with a fan-in and fan-out of one. This implies that an appropriate topology for the interconnect is a ring. The large bandwidth on and off the chip made available by the optics implies that only a small fraction of that bandwidth need be made available to each node in the ring.

4.1 Ring Topology

Figure 2 illustrates the logical topology of a multiring [10]. The physical ring is divided into channels, one channel per node on the ring. Each channel is associated with a particular destination node, and this destination node receives messages from the other nodes on the channel. In the 4 node example of Figure 2, the outside ring is associated with node 1, the next-to-outside ring is associated with node 2, the next-to-inside ring is associated with node 3, and the inside ring is associated with node 4. In the design of [10], the physical links are constructed using optical fiber and exploit tunable lasers to implement WDM multiplexing. Note that tunable lasers are not necessary in this system since there are sufficient channels using space division alone.

With the multiring topology, each channel can be thought of as a daisy chain terminating at the destination. The implementation of a four node (N1, N2, N3, and N4) multiring using an optical interconnect is illustrated in Figure 3. The multiring has the following advantages:

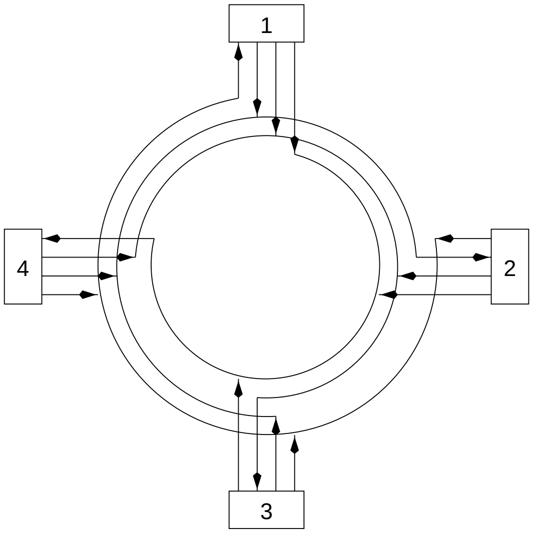

Figure 2: Multiring topology.

- Ideally Suited for Free Space Optical Interconnection: The optical fan-in and fan-out of each node is one. Single-hop communication is only with the two nearest neighbors.

- No Need for Explicit Destination Address Specification: An incoming message landing on the detectors assigned to channel i on node i's receiver automatically indicates that the message destination is node i.

- No Need for Explicit Routing: Since each channel is associated with a singe receiver node, there is no complex routing necessary. If the node receiving the message is not the destination node, only a fixed forwarding operation is performed.

In the example, if node 4 wants to send a message to node 2, it will send the message on channel 2. The message will first be received on channel 2 of node 1's detector array. Node 1 will then repeat the message on channel 2 of its VCSEL array. The message is then directed to channel 2 of node 2's detector array and is thus delivered to node 2.

Note that in both Figures 2 and 3, the number of signal paths between each node is not four, but three. This is due to the fact that node i need never send messages to itself via the optical interconnect, and does not need an outbound optical path. In general, $N-1$ optical channels are required between any pair of nodes, implying that the number of VCSEL-detector pairs that can be allocated to each channel is $\lfloor M^2/(N-1) \rfloor$.

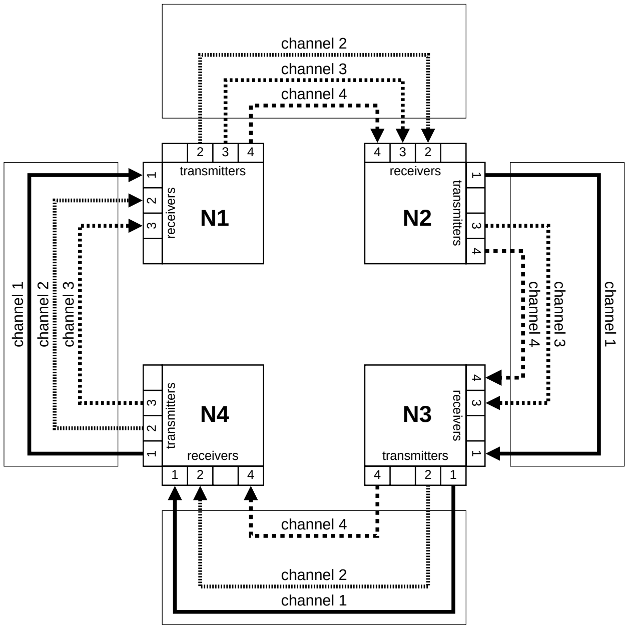

Figure 3: Conceptual diagram of a 4 node multiring.

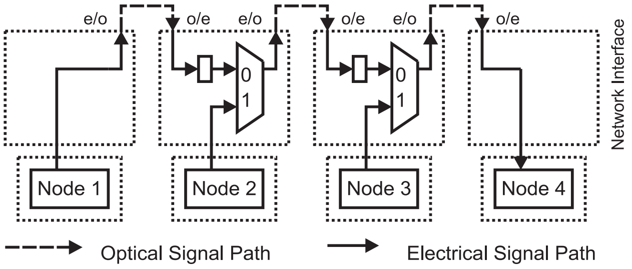

Figure 4: Channel 4 of a 4 node ring.

4.2 Network Interface/Memory Interface

Figure 4 illustrates an individual channel (channel 4) in a 4 node ring. At the top of the figure, the optical signal path is shown as a set of dashed lines. Electrical-to-optical (e/o) and optical-to-electrical (o/e) signal conversions are shown as vertical transitions out of and into the network interface associated with each node. Node 1, as the first node in the channel, requires only an e/o conversion in the network interface. Node 4, as the destination node, requires only an o/e conversion. The remaining nodes have both an o/e and e/o conversion as well as a buffer and selection multiplexer. The multiplexer is used to grant access to the outbound optical link, either from the upstream neighbor or the local node.

An important consideration in the system is how to manage the selection multiplexers in each network interface. In [2], consideration is given to both connection-oriented approaches as well as packet-based approaches. Here, we assume that message delivery is cell-based (i.e., messages are segmented into fixed length cells for delivery across the optical interconnect and reassembled at the destination), and that at cell boundaries the selection multiplexers give priority to upstream traffic whenever it is present. The primary benefit of this design decision is that the buffers in the network interface need be only one cell size in length. This is important, since at the optical data rates available, buffer requirements can quickly become very large. The disadvantage of this choice is that there is an inherent bias in media (i.e., channel) access given to the upstream nodes. A technique to address this inherent bias is described in Section 6.

Extremely high bandwidth in internode communications does not improve overall system performance if there is a bottleneck at the nodes. Unfortunately, the current bus-based I/O bandwidth of even high performance processors is generally unable to keep up with the data rates available in the optical interconnect. We address this issue by changing the standard memory system design and providing a direct path from the network interface into the node's memory, bypassing the I/O bus completely. A block diagram of an individual node is shown in Figure 5.

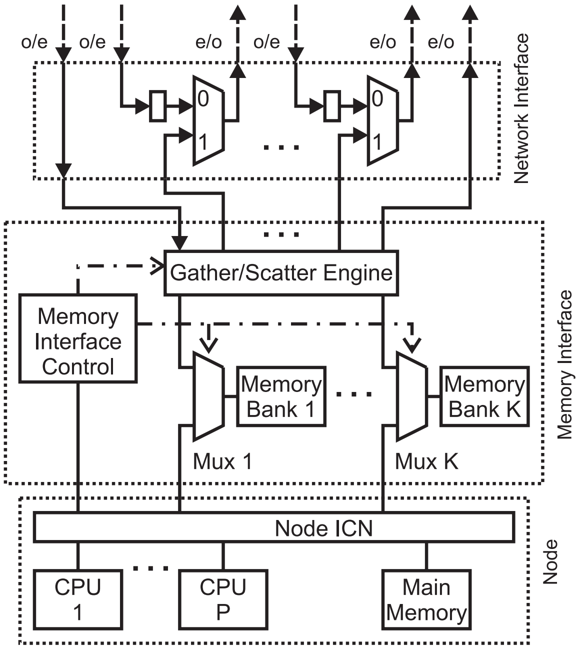

Figure 5: Node architecture.

The network interface at the top of the figure shows the signal conversions, forwarding cell buffers, and selection multiplexers for all of the channels in the multiring. Note that there will be a single incoming channel (the channel which has the local node as its destination) shown on the left of the figure, a single outbound channel that does not require a selection multiplexer shown on right of the figure, and $N-2$ pass-through channels.

To provide a fast path for local data (i.e., passing into or out of a node), the traditional main memory-bus (or general ICN) datapath is augmented with dual-port memory modules which can be shared both by the local bus and the optical interconnect. Muxes 1 through K act to select either the local 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 many embedded applications.

The prototypical use of the gather/scatter engine is to support corner turn operations on STAP applications [8]. Here, the data is commonly block-decomposed matrices, and the communication required from one node 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 node 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.

5 Interconnect Performance

In order to investigate the performance of the optical ring interconnect, a discrete-event simulation model has been developed. This simulation model was implemented within the ICNS framework [4] using the MODSIM III language. We model a single channel of the multiring initally at the level of abstraction illustrated in Figure 4. This will be refined later in Section 6.

The initial performance investigations center around the impact of the design choice associated with the selection multiplexers present in the network interface of Figure 4. Recall that to minimize buffering requirements within the network interface, upstream nodes are given priority over locally generated traffic for access to a channel on the ring. This has the potential to generate significant unfairness in the system. To quantify the performance implications of these decisions, one channel (channel eight) of an eight node interconnect was simulated with uniform traffic. Each of the seven source nodes generated messages from a Poisson arrival process, and message lengths were exponentially distributed with a mean of 100 cells. Defining the message latency as the time from initial message generation at the source to the time that the last cell of the message is delivered to the destination, the mean message latency for each source was measured as a function of offered load.

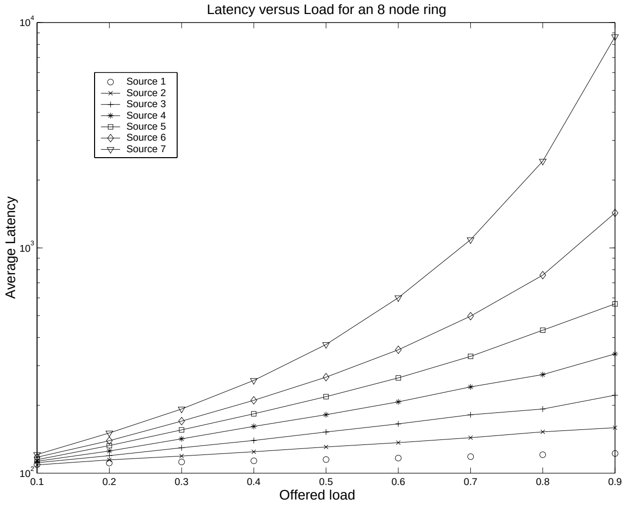

Figure 6: Message latency vs. offered load.

The results of this initial simulation are shown in Figure 6. The impact of giving priority to upstream nodes is clearly significant, especially at high load. The highest priority node (node 1) experiences latency near the minimum (100 cell times) independent of load, while the lower priority nodes experience significantly increasing latencies with increasing load.

The impact of priority is even more dramatic when throughput is considered under overloaded conditions. Figure 7 plots equal-time snapshots of the number of cells sent from each source under overloaded conditions (mean message size of 256 cells and an offered load of 1.5). The horizontally connected points represent snapshots at a given point in time of the number of cells sent from each source. The snapshot interval is 1000 cell times. As is clear from the graph, the bandwidth consumed by the high priority nodes results in lower throughput (even starvation) for the low priority nodes. In this example, node 5 has a throughput of approximately half that of nodes 1 through 4, while nodes 6 and 7 are completely starved.

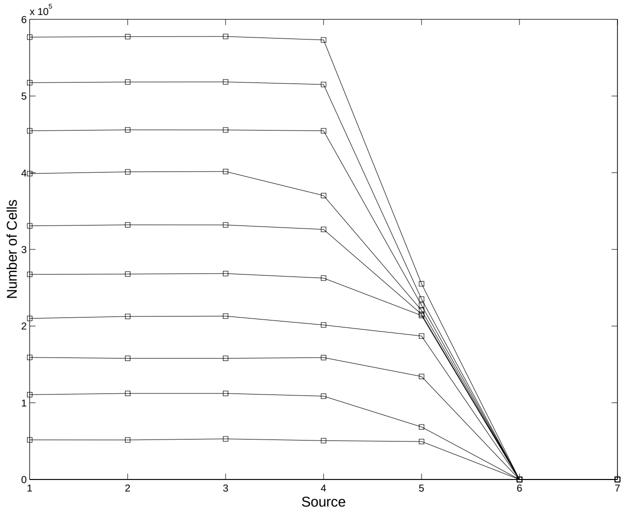

Figure 7: Throughput vs. source, offered load = 1.5.

The degree of unfairness experienced by nodes on the ring is clearly an unacceptable circumstance for many applications. In the following section, we describe an approach to deal with this issue.

6 System Issues

The above optical interconnect has two important issues that must be addressed before it can be effective: reliability and fairness. The reliability of this type of system is assessed in [2]. Here we address the issue of fairness.

The performance results in Section 5 illustrate the need for a more effective mechanism to arbitrate access to individual channels on the multiring. While it is often advantageous for an arbiter to allow uniform access to all the nodes contending for a channel, this may be application dependent. For example, it may be desirable to give a subset of the nodes more access to a channel (i.e., more bandwidth) than other nodes. An approach that suits the multiring is the Deficit Round Robin (DRR) scheduler.

DRR scheduling was introduced in [13] for use in internet switches and routers, where the contention is for an output link. It was modified in [3] for use in a Banyan topology interconnection network, and is briefly described in [2] for use in a multiring. The DRR scheduler has the following attractive properties:

- Flexibility. Nodes can be given different amounts of access to a channel by tuning parameters built into the protocol.
- Fast Decision Making. The DRR algorithm is fast since it needs to only examine the node in question to decide whether it should be given access to the channel.
- Fairness. DRR has been proven fair to the following extent: at any time, for equal priority channels, the difference in the amount of access granted to the most advantaged contender and the most disadvantaged contender is no more than three times the maximum message size.

Here, we describe the DRR scheduler as adapted for a multiring topology and present performance results for the interconnect with the DRR scheduler present.

In the original development of the DRR scheduler, all the information necessary to make scheduling decisions was present at a common location. This is not true for the multiring, and the original mechanisms must be adapted to work in this environment. Since every channel is associated with a single destination, we assign the DRR scheduler for that channel to its associated destination. Prior to sending a message on channel j, a node sends a control signal to node j requesting access to the channel. When the DRR scheduling algorithm (executing on node j) decides that sender i should have access, it replies with a control signal to i granting access to the channel.

The DRR scheduling algorithm, executing at each destination, maintains $N-1$ deficit counters, one for each potential message source. Each source node i is also assigned a quota q_i, indicating its relative bandwidth assignment on the channel. If all the quotas are equal, $q_i = q, \forall i$, the scheduler is to give equal access to the channel to all source nodes.

Upon receipt of a control signal from node i requesting access, the scheduler compares the size of the request to node i's deficit counter. If the request is to be granted (the message size is less than the deficit counter), a grant control signal is sent to node i and i's deficit counter is reduced by the size of the message. If the request is not granted (the message size is greater than the deficit counter), the control message remains in a request queue and is reconsidered in the next round (defined below).

Once per round, the deficit counters associated with each source are increased by their quota q_i. A round is defined as a period during which each source node contending for access is given the total allowed access as defined by its deficit counter. That is, a round is complete when every source is either not contending for access to the channel or has a deficit counter less than the size of its pending message. Details of the DRR scheduler are given in [13], and a complete description of its adaption to the multiring topology, including the implementation of in-band control signal delivery, is presented in [9].

In order to investigate the effects of DRR on the performance of the interconnect, the discrete-event simulation model was expanded to include the DRR functionality. This includes the implementation of an in-band control signal delivery mechanism, the use of this mechanism to deliver the control signals required by the DRR protocol, and the execution of the DRR algorithm [9].

Figure 8 illustrates the impact of the DRR scheduler in an overload situation. As before (in Figure 7), we are simulating a total offered load of 1.5 for 7 sources all contending to send data to node 8. The snapshot interval is identical to that of the previous graph. The only distinction between the two simulations is the presence of a DRR scheduler.

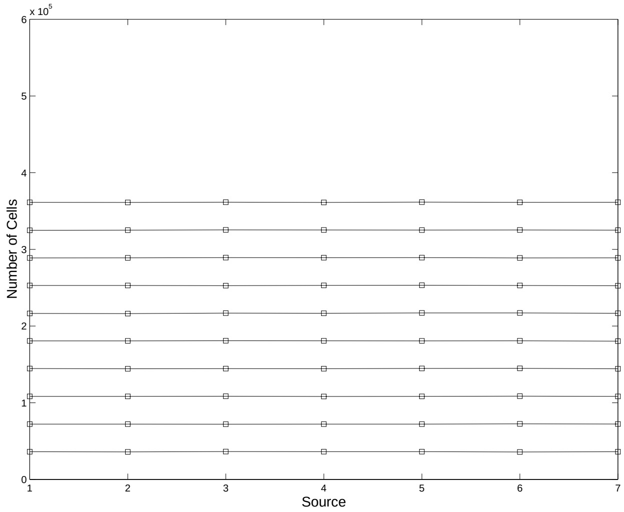

Figure 8: Throughput vs. source with offered load = 1.5 and DRR scheduler.

Unlike the earlier case, where downstream nodes (5, 6, and 7) received limited (or no) access to the channel, here each of the source nodes receives uniform access. The effective bandwidth delivered to each source is lowered (reflecting the total bandwidth capability of the channel), and it is fairly allocated to the contending sources.

Figures 9 and 10 show the impact of a DRR scheduler on the latency performance of the interconnection network. Figure 9 is similar to Figure 6 with some additional information. The graph is in a bar-and-whisker format, with the bars representing the mean latency measured at nodes 1, 4, and 7 and the vertical line, or whisker, representing the standard deviation in the latency experienced at that node. Figure 10 represents the same measurements taken on a system that includes a DRR scheduler.

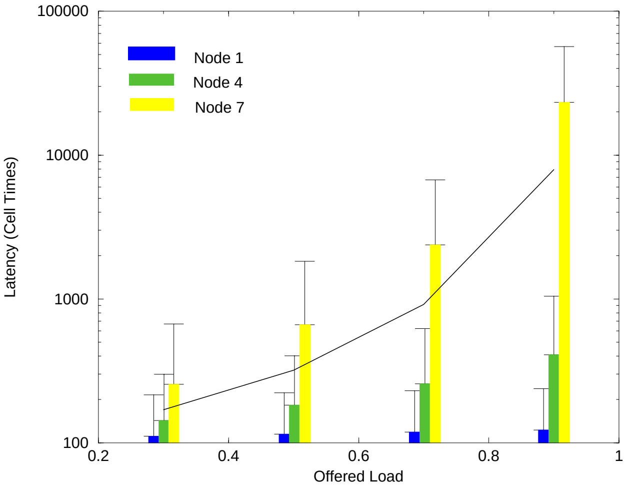

Figure 9: Mean latency and standard deviation without DRR. Solid line is average across sources.

The impact of the DRR scheduler is readily apparent in the two figures. In Figure 9 (without DRR), there is a significant variation between the latencies experienced by the different nodes in the ring. In Figure 10 (with DRR), these latencies have been evened out. To examine the impact on overall latency, the solid line plots the average latency (across all 7 source nodes) without the DRR scheduler. (The solid line is plotted on both graphs for comparison purposes.) On Figure 10, the dashed line plots the average latency (again across all nodes) with the DRR scheduler. Note that average latency has not grown with the introduction of DRR (the small decrease in latency is not statistically significant). DRR thus significantly improves performance by equalizing the latencies, without increasing the average latency over all sources.

7 Summary and Conclusions

This paper has presented an architectual design and performance analysis of a photonic ring interconnect intended for use in embedded multicomputers. The target applications are ones in which interprocessor data communication throughput is an important performance consideration. With near term technology, terabits/sec of bandwidth is achievable.

The physical ring topology is logically organized as a multiring, with each channel of the multiring associated with a distinct destination node. Both a network interface design and a memory interface design are described, with the goal of matching the bandwidth of the local node to the network thus ensuring that it

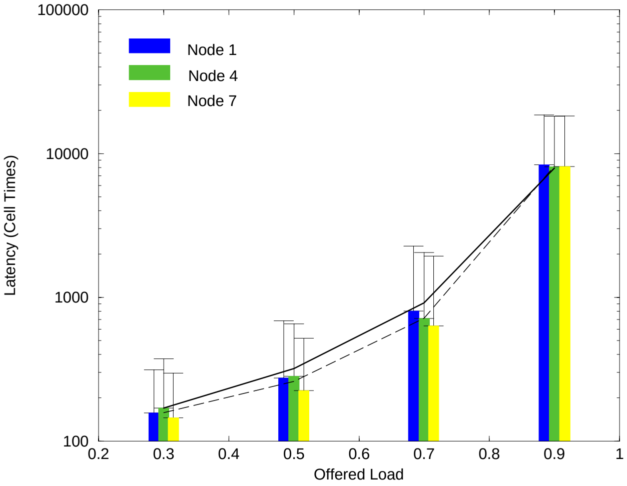

Figure 10: Mean latency and standard deviation with DRR. Solid line is average across sources without DRR and dashed line is average across sources with DRR.

does not become an application bottleneck.

In addition to the basic architecture, we present performance results quantifying the unfairness introduced in the system by low-level design choices. The use of a deficit round robin scheduling protocol is proposed to address the inherent unfairness in the system, and the performance is considered with the scheduling protocol in place.

At present, the system described in this paper is not yet implemented. The performance results presented here are from a discrete-event simulation model. A number of the constituent pieces, however, have been constructed. A 16 × 16 array of VCSELs and detectors as well as both rigid and flexible light pipes have all been demonstrated (by our partners in DARPA's VLSI Photonics Program), and larger arrays are planned. This is a viable technology that is available now and is likely to have a significant impact on future multicomputer designs.

References

[1] M. Chateauneuf et al. Design, implementation and characterization of a 2-D bi-directional free-space optical link. In *Proc. of Optics in Computing*, pages 530–538, June 2000.

[2] Ch'ng Shi Baw, R.D. Chamberlain, and M.A. Franklin. Design of an interconnection network using VLSI photonics and free-space optical technologies. In *Proc. of 6th Int'l Conf. on Parallel Interconnects*, pages 52–61, October 1999.

[3] Ch'ng Shi Baw, R.D. Chamberlain, and M.A. Franklin. Fair scheduling in an optical interconnection network. In *Proc. of MASCOTS*, 1999.

[4] 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.

[5] Mark Franklin, Abhijit Mahajan, and R. Martin Arthur. Parallel implementations of an ultrasonic image generation algorithm using MPI. In *Proceedings of the Parallel and Distributed Computing and Systems Conference*, November 1999.

[6] H. Kosaka et al. A two-dimensional optical parallel transmission using a vertical-cavity surface emitting laser array module and an image fiber. *IEEE Photon. Tech. Lett.*, 9:253–255, 1997.

[7] Y. Li, E. Towe, and M. Haney, eds. *Proc. on Short Distance Optical Interconnections in Digital Systems.* IEEE, 2000.

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

[9] Abhijit Mahajan. Performance analysis of an optical interconnection network. Master's thesis, Washington University, Saint Louis, MO, 2000.

[10] M. Marsan et al. All-optical WDM multi-rings with differentiated QoS. *IEEE Communications Magazine*, pages 58–66, February 1999.

[11] J.A. O'Sullivan, M.A. Franklin, M.D. DeVore, and R.D. Chamberlain. Analysis of computational system performance in automatic target recognition. In *Proc. of High Performance Embedded Computing Workshop*, September 2000.

[12] D. Plant et al. A 256 channel bi-directional optical interconnect using VCSELs and photodiodes on CMOS. In *Proc. of Optics in Computing*, pages 1046–1054, June 2000.

[13] M. Shreedhar and G. Varghese. Efficient fair queueing using deficit round robin. In *Proc. of SIGCOMM*, pages 231–243, August 1995.