Optical Network Reconfiguration for Signal Processing Applications

Roger D. Chamberlain
Mark Franklin
Praveen Krishnamurthy

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

Optical Network Reconfiguration
for Signal Processing Applications

Roger Chamberlain, Mark Franklin, and Praveen Krishnamurthy
Computer and Communications Research Center
Washington University, St. Louis, Missouri
{roger,jbf,praveen}@ccrc.wustl.edu

Abstract

This paper considers a class of embedded signal processing applications. To achieve real-time performance these applications must be executed on a parallel processor. The paper focuses on the multiring optical interconnection network used in the system and specifically on the performance gains associated with utilizing the bandwidth reconfiguration capabilities associated with the network. The network is capable of being reconfigured to provide designated bandwidths to different source-destination connections both across rings and within a ring. The applications each consist of a sequence of alternating communication and computation phases. The sequence continues until execution of the application is complete. The effect of reconfiguration on application performance is explored using simulation techniques. The results indicate that substantial performance gains (speedups of 2 or more) can be achieved for this application class.

1   Introduction

As the cost of computers has decreased and their performance increased, the use of parallel processors in embedded real-time systems has grown. Associated with this growth has been a need for higher processor interconnect bandwidth. Recent advances in VLSI photonic technology have led to the development of photonic communication technologies to satisfy this need. This paper considers a specific class of real-time applications associated with signal processing (e.g., synthetic aperature radar image formation, beamforming).

Successful design and implementation of optical interconnects with terabit per second bandwidth capacity has been demonstrated in a number of projects [5, 10]. Of interest here is a multiring architecture that utilizes optical links between VLSI chips. The individual rings which make up the multiring are each associated with a given destination processor, with one ring for each destination. The overall organization has been described in [2] and is reviewed in Section 3. The design is targeted at multicomputer systems and associated applications requiring frequent, massive data transfer among processors. Examples of this include embedded real-time signal processing applications.

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 [10], and designs with $M = 32$ (currently under construction) have theoretical raw bandwidth greater than 1 Tb/s.

Applications implemented on such a multiprocessor typically consist of a sequence of compute-communicate phases with each having different source-destination communication bandwidth requirements. Overall performance can be improved if there is the ability to reconfigure the interconnection network to match application bandwidth requirements associated with each phase. Within the multiring architecture, reconfiguration is achieved by exploiting two methods:

- **Deficit Round Robin (DRR) Algorithm:** Within a given ring (associated with a single destination processor) use of DRR permits flexible control of the bandwidth allocation from each source to the designated destination [2].
- **Laser Channel Allocation (LCA) Algorithm:** The number of optical paths, and hence the bandwidth, associated with each ring can be changed. This effectively determines the aggregate individual ring bandwidth allowable between all sources and a given destination.

As shown later in this paper, the effect of proper reconfiguration is a significant reduction in the time associated with data communication, and thus with overall execution time. This paper discusses the two reconfiguration methods introduced above, and the performance gains achievable when employing them on a selected application set.

Section 2 is an introduction to the optical technology used in the interconnect architecture. Section 3 presents the architecture details of the interconnect and also describes the concept of reconfigurability as applied in this case. Section 4 describes the application class that was simulated. Section 5 describes performance results from a discrete-event simulation model of the system. Section 6 summarizes our results and concludes the paper.

2 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 on a single chip [10]. In this system, the VCSELs arrays and photodiode arrays were flip-chip bonded to a CMOS chip using heterogeneous integration techniques. This is illustrated in Figure 1 which shows two separate side-by-side 2×2 VCSEL and detector arrays bonded to a CMOS chip. The VCSEL and detector arrays provide for communication, while the CMOS chip would provide for processing and, in our design would contain a full processor.

While the demonstration of [10] used bulk optics to deliver light between ICs, designs have been investigated utilizing both rigid optical links [3] optimized to be misalignment tolerant (useful for chip-to-chip links on a board), and flexible fiber imaging guides [6] (useful for board-to-board links). Given the vertical nature of the VCSEL process, both approaches require connection to the top of the arrays.

The availability of a large number of VCSEL/detector pairs in the optical interconnect suggests the partitioning of the optical links into sets with each set being associated with an individual channel (i.e., space-division multiplexing). Figure 2 illustrates the allocation of VCSELs and detectors for a four channel system utilizing 16×16 arrays of optical elements. As shown in the top of the figure, one quarter of the elements are used for each channel. Each square in the top view of Figure 2 contains a single VCSEL or detector.
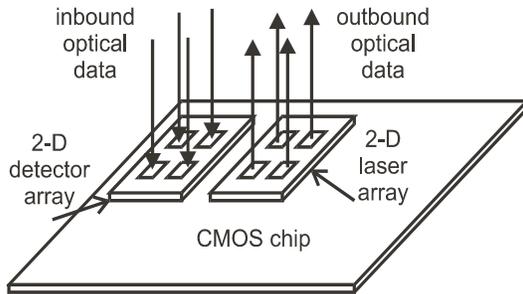
Figure 1: Optical technology illustration. VCSEL/detector arrays are bonded to the CMOS chip. Inbound data is received and outbound data is delivered vertically.
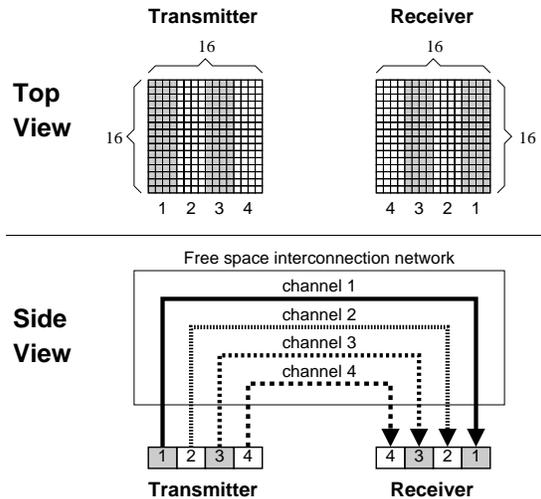

Figure 2: Allocation of VCSEL/detector pairs to a four channel system. 16×16 VCSEL/detector arrays are used, with a 4×16 array allocated to each channel.

If the individual element communicates at 1 Gb/s, this yields $16^2/4 = 64$ Gb/s per channel. The side view of the figure illustrates (conceptually) how two adjacent chips might communicate.

3   System architecture

3.1   Multiring topology

The overall system considered here is a multicomputer that is embedded into a larger application specific system. The final design decisions (dimensioning, configuring, etc.) are guided by the specific application being implemented.

The multicomputer consists of N nodes, with each node consisting of a symmetric multiprocessor containing P processors and local memory. Some fraction of a node's local memory is designated as *communication memory* and is the primary data interface to the optical interconnection network. 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.

The free space optical technologies described here are most cost-effective when used with a fan-in and fan-out of one and a topology meeting this fan-in/fan-out goal is a ring. While there are many approaches to developing a ring based interconnect, given the very high bandwidths available, the multiring [9] design of Figure 3 has been chosen.

In the 4-node example of Figure 3, each of the four rings is associated with a given destination node. The outside ring, for example, is associated with node 1 and the next-to-outside ring is associated with node 2. The inside ring is associated with node 4. With the multiring topology, each ring can be thought of as a daisy chain terminating at the destination node. Thus, communication between node i and node j requires that node i send its message on the ring that has node j as the destination.

The physical implementation of the multiring requires that the VCSEL/detector array be divided into channels (as indicated earlier) and that each channel correspond to a given

ring (and thus destination node). This destination node receives messages from the other nodes on the channel. Note that while even higher bandwidth can be obtained using tunable lasers to implement WDM multiplexing [9], in the system considered here there is sufficient bandwidth available using space division multiplexing alone that this is not considered.
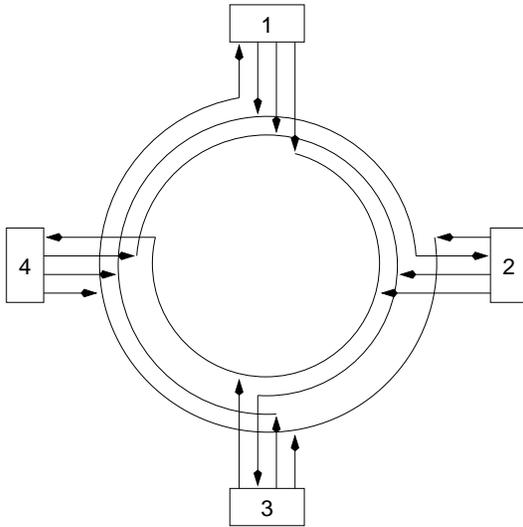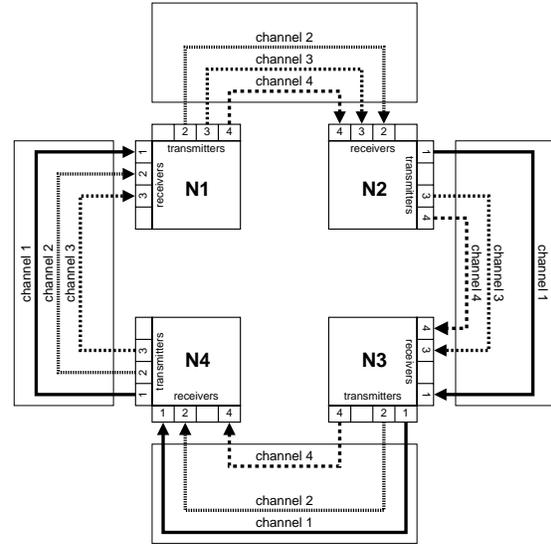

Figure 3: 4 node multiring.

Figure 4: 4 node multiring implementation.

The implementation of a four node (N1, N2, N3, and N4) multiring is illustrated in Figure 4. The multiring has the following advantages:

- 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 need be 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 therefore directed to channel 2 of node 2's detector array and is thus delivered to node 2.

Note that in both Figures 3 and 4, 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 allocated to each channel (assuming uniform allocation) is $\lfloor M^2/(N-1) \rfloor$. The uniform allocation assumption is removed below to support variable bandwidth allocation across rings.

3.2 Reconfigurability

Reconfigurability in this system refers to the allocation of interconnect bandwidth resources based on application requirements. As shown above, the multiring architecture consists of individual rings with each ring assigned to a given destination node. As indicated, with the LCA algorithm, the bandwidth of each of the rings can be set by allocating the number of VCSEL/detector pairs in a given channel (and thus its associated ring). This technique can be used to allocate bandwidth on the basis of destination nodes. Thus some rings may be assigned increased bandwidth while others may have less bandwidth. In this way the bandwidth associated with traffic to the nodes can be varied and matched to the needs of the application. The assignment itself is possible because the information received on a detector and transmitted from a VCSEL passes through the CMOS chip (see Figure 1) where routing logic can dynamically assign interconnect paths to different rings.

Within a ring, media access is arbitrated using the DRR fairness protocol [2, 11]. The protocol supports the assignment of arbitrary bandwidth ratios to sources in a ring, accomplished by varying the "quanta" allocated to each of the sources. Over a given time period, each source has a quanta (i.e., equivalent to a bandwidth resource) that it can utilize. The DRR protocol determines whether a source has used all its quanta and determines interconnect access so that each source obtains its minimum quanta over a the specified time period. Thus, within a ring, the bandwidth associated with each of the sources can be matched to the needs of the application. Note that with DRR, unused quanta are reallocated to sources requesting bandwidth in a round robin fashion.

The above two-level mechanism ensures that the interconnect can be configured for any arbitrary bandwidth allocation between source-destination pairs (flows). Naturally, this is constrained by the total available interconnect bandwidth.

We have simulated the performance of the optical interconnect with reconfiguration using a number of applications. In these simulations we assume that there is *a priori* knowledge of the per flow communication bandwidth required for each application phase. Reconfiguration across rings (LCA), and within a ring (DRR), is changed before the start of each phase based on this *a priori* knowledge. This knowledge is generally available for the set of applications considered (e.g., synthetic aperture radar, beamforming).

In contrast with static reconfiguration, with dynamic reconfiguration there is no prior knowledge of the application phase bandwidth requirements. In this case a learning algorithm would be used to determine application requirements during execution and, based on this information, perform the appropriate reconfiguration. Determining bandwidth requirements dynamically is analogous to determining the working set in virtual memory systems, however, in this situation we would need to determine the "working bandwidth." Just how to do this is a current topic of research and is not considered here.

By performing optimal bandwidth allocations based on knowledge of the application requirements, the static allocation realizes the best performance possible and thus represents an upper bound on dynamic allocation performance. The next section considers the applications and application models used.

4 Applications and application models

Even without reconfiguration, utilization of the optical interconnect will result in significant performance gains for those applications where, in conventional systems, communication is a bottleneck. For example, in systems executing iterative algorithms, if the ratio of communication time to compute time is greater than one, using the sort of optical

interconnection network described above will generally result in considerable performance gains. Reconfiguration further enhances these performance gains.

For the class of applications considered, a cycle consists of a computation phase followed by a communication phase. A communication phase is defined in terms of two components:

- **Communications Pattern**: The set of source-destination pairs that will communicate during the phase.
- **Communication Volume**: The amount of information that needs to be transferred between each source-destination pair.

Both the pattern and volume requirements for an application may vary from one phase to the next. Computation requirements are included as a variable. With the computation set to zero, we explore the reconfiguration gains associated only with the communication component of application performance. With other values we explore the reconfiguration gains associated with overall application execution times. For all our simulations, the number of nodes was set at eight.

Two sets of applications are considered. The first consists of two real applications where, from an understanding of the application, the communication patterns and volumes are known. The second set consists of synthetic applications whose properties have been chosen randomly from a set of common communications patterns.

The two real applications include synthetic aperture radar (SAR) image formation, and beamforming (BF). The SAR application, for example, can be viewed in terms of the phases shown in Figure 5. For this application, the first communication phase consists of data being input from the sensor array (a broadcast). The first computation phase consists of range processing. The second communication phase is a corner turn operation (an all-to-all pattern). The second computation phase is azimuth processing, and the final communication phase is the output of formulated SAR images (a reduction).


Figure 5: SAR phases.

The communication patterns associated with each SAR phase are shown in Figure 6. Nodes 1 and 8 correspond to the input and output nodes respectively. Nodes 2 through 7 correspond to processor nodes which perform the computation. Thus, in phase 1 the communication pattern corresponds to distributing the input data from node 1 to processing nodes 2 through 7. In phase 2, an all-to-all exchange of data between the processing nodes takes place, while in phase 3 a reduce operation occurs which aggregates the final image from the processor nodes to the output node 8. Similarly, the properties of the BF application have been determined and modeled.

Ten synthetic applications were also analyzed. For each, the following three application parameters were generated in a random fashion:
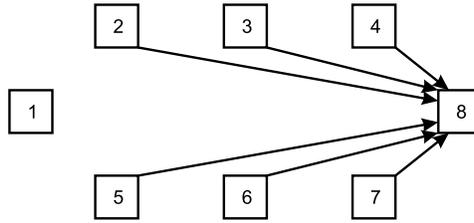
- **Number of Phases:** Uniformly distributed between 3 and 6.
- **Communications Pattern:** Four communications patterns commonly associated with space-time adaptive algorithms were considered:
    - **All-to-All:** All the nodes exchange data with each other.
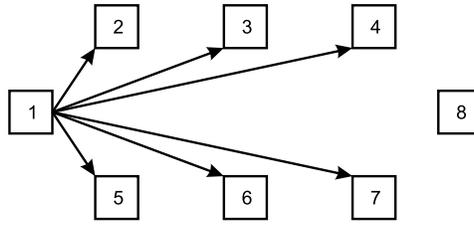
(a) Broadcast from sensor node to compute nodes.
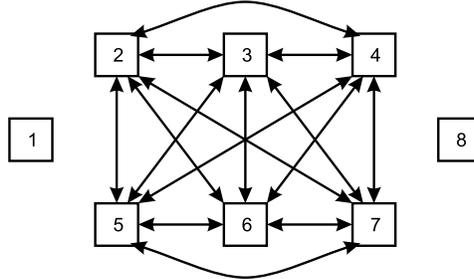
(b) Corner turn between compute nodes.

(c) Reduction from compute nodes to output node.

Figure 6: SAR communication patterns.

- **Broadcast:** One randomly selected node sends information to a random selection of other nodes.
- **Reduce:** A random selection of nodes sends information to a single randomly selected destination.
- **Point-to-Point:** A random set of source-destination pairs are selected.

• **Communication Volume:** For each flow associated with each pattern, the amount of information to be transferred was randomly selected from a fixed set of message sizes that spanned two orders of magnitude.

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 [1] using the MODSIM III language. Earlier work using this simulator concentrated on investigations relating to multiring design [4], the effectiveness of the DRR protocol [2], and the performance of the multiring with selected applications [8].

As indicated earlier, two levels of reconfiguration are available; one based on DRR (within a given ring), and one based on LCA (across the entire multiring). To separate out

the performance effects of each of these reconfiguration methods four sets of simulation experiments were performed for each application:

- **UA - Uniform Allocation:** Over all phases, the bandwidth was divided evenly among the rings and, within each ring, sources were given equal quanta. This ensures uniform allocation over all source-destination pairs and represents the base case where no reconfiguration is done.
- **DRR - DRR Quanta Allocation:** Available bandwidth is evenly divided among the rings. Within a ring, knowing the bandwidth requirements of each source-destination pair (or flow), the quanta associated with pairs in the ring are adjusted to reflect the application flow bandwidth demands. This is done at the start of each phase and represents ring-level reconfiguration.
- **LCA - Laser Channel Allocation (LCA):** Knowing the bandwidth requirements of each source-destination pair (or flow) one can determine the bandwidth requirements associated with each ring. Based on this, LCA divides up the total bandwidth available to reflect the bandwidth needs of each ring. This is done at the start of each phase. Within each ring, the quanta associated with each flow are set equal.
- **DRR-LCA - DRR and LCA Together:** Knowing the bandwidth requirements of each source-destination pair, both DRR quanta and LCA ring bandwidth allocations are performed at the start of each phase.

By performing each of these simulation experiments, the performance effects of each of the reconfiguration methods can be determined.

The performance measures of interest include maximum and mean completion times, and variability in completion times. The maximum and mean completion times (across flows within an individual communication phase) are of interest both in absolute terms and as a speedup relative to the uniform bandwidth allocation. The variability in completion times is shown using the coefficient of variance for the completion times associated with each flow. This variability measurement is an indication of the fairness of the system. Values near zero imply equal delivery times, while values approaching (or exceeding) one indicate variability of the same order as the mean completion time.

The only communication pattern that requests variable communications volume across the message set is the point-to-point pattern. As a result, it is the only pattern that will have meaningful results for evaluating the effectiveness of the fairness provided by the DRR protocol within a ring. For this reason, the first set of performance data is restricted to the 14 communication phases (extracted from all 12 applications) that correspond to a point-to-point pattern.

Since the DRR protocol utilizes unrequested bandwidth allocated to one source for another source that makes a request, the maximum completion time for a phase should not change with DRR reconfiguration. The variability in completion times should, however, significantly improve. LCA reconfiguration, on the other hand, should impact both maximum completion time and variability.

Figures 7 and 8 show the maximum and mean completion times and completion time variability for all 14 point-to-point communication phases. In each plot, the performance is given for the initial uniform allocation, reconfiguration within a ring (via DRR), reconfiguration across rings (LCA), and the combination of both reconfigurations (DRR-LCA). The completion times are in cell times, which is the base time unit for the simulation.
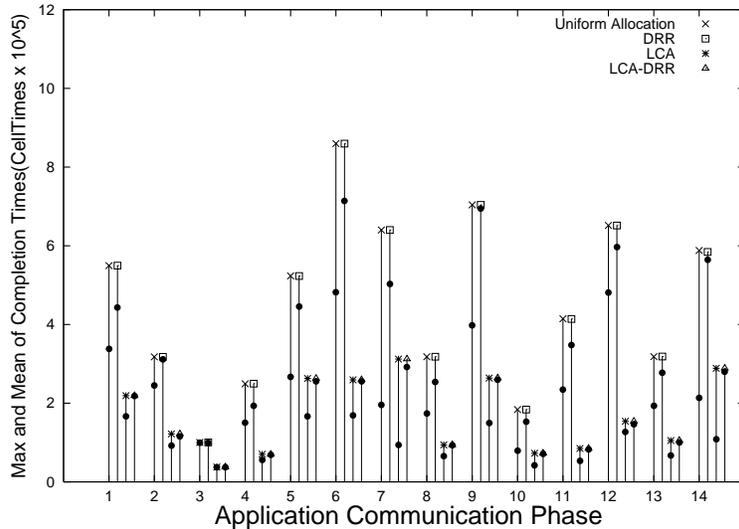
Figure 7: Maximum and mean ($\bullet$) completion times for point-to-point comm. phases.

Observing both maximum and mean completion times and variability for each phase, we observe that the DRR reconfiguration has no effect on maximum completion time but a significant impact on both the mean completion time and variability. This clearly illustrates the tradeoff associated with using DRR. That is, use of DRR assures greater fairness in bandwidth allocation, and thus a reduction in variability, however, at the cost of increasing the mean of the completion time. The LCA reconfiguration on its own has a limited impact on variability and but a dramatic impact on maximum completion time. When the two reconfiguration mechanisms are combined, we see the combined benefit of a significant decrease in maximum completion time and variability reduced to near zero.

The remaining results include data from all communication phases of all applications. Figure 9 shows the improvement in maximum completion time for the communication phases of each application with reconfiguration. In each case, the decrease in communication overhead associated with the application is significant.

Though all the applications were restricted to 3 to 6 phases, there is appreciable variation in the communications performance improvement achieved across the applications. The range of speedups obtained was from 1.9 to 7.1. The average speedup across all the applications was approximately 4. The large variation in the speedup of particular applications can be partially attributed to the type of communication pattern predominant in that particular application. Figure 10 shows the maximum, median, and minimum speedup that each communication pattern yielded across the set of applications.

The performance values presented up to this point have been exclusively concerned with the communication phases of the application. To explore the overall performance impact, one must integrate the communications performance into a performance model for the application as a whole. Since computation and communication are non-overlapping in the applications of interest, one can use Amdahl's Law to relate the performance improvement in the communication phases to the performance improvement of the complete application. Here,

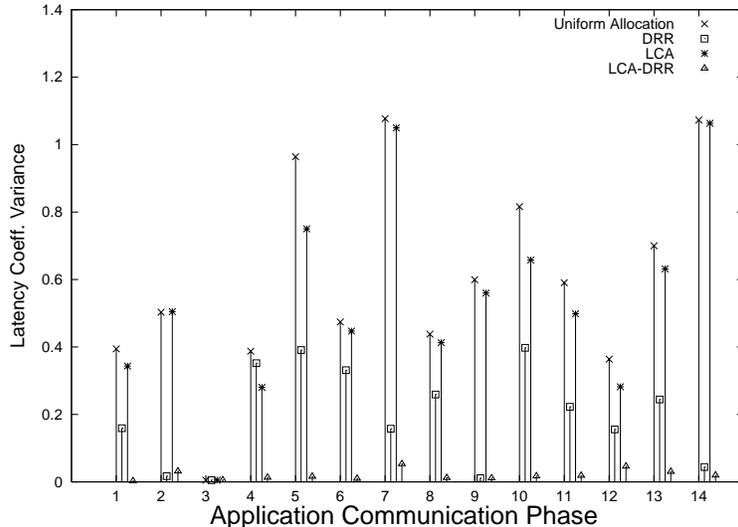$$Speedup_{overall} = \frac{1}{f_{comp} + \frac{f_{comm}}{Speedup_{comm}}}$$

Figure 8: Variability in completion times for point-to-point communication phases.

where f_{comp} is the fraction of the original execution time associated with computation and f_{comm} is the fraction of the original execution time associated with communications.

Figure 11 plots the application speedup versus the ratio of communication time to computation time in the original (uniform allocation) configuration. The communicate to compute ratio is plotted from 0.5 (communication takes one half the time as computation) to 2.0 (communication requires twice the time as computation). This range is typical of the applications of interest. The three curves represent the performance improvement associated with the minimum, mean, and maximum increase in communications completion time across the applications.

The immediate conclusion one can draw from this graph is not new, but has been well understood for a long time. That is the fact that the overall performance gain attributable to an enhancement associated with only a portion of the original execution time is limited. None the less, the overall speedups associated with reconfiguring the communications infrastructure (the optical interconnect) are significant. A 20% performance gain is predicted even under fairly pessimistic assumptions, and a potential doubling of performance is possible.

6 Summary and conclusions

This paper has presented the performance gains achievable in a class of embedded signal processing applications through the use of reconfiguration in an optical interconnection network. The multiring architecture of the optical interconnection network was described, and two distinct reconfiguration mechanisms were presented. Within a ring, the DRR fairness protocol allocates instantaneous bandwidth across the sources contending for an individual destination. If some sources do not utilize their allocated bandwidth, the excess bandwidth is then distributed across the contending sources. Across the multiring, the LCA mechanism supports the flexible assignment of bandwidth to each ring (and its associated destination).

The performance implications of this reconfigurability were presented using two real applications and ten synthetic applications, all representative of a class of signal processing
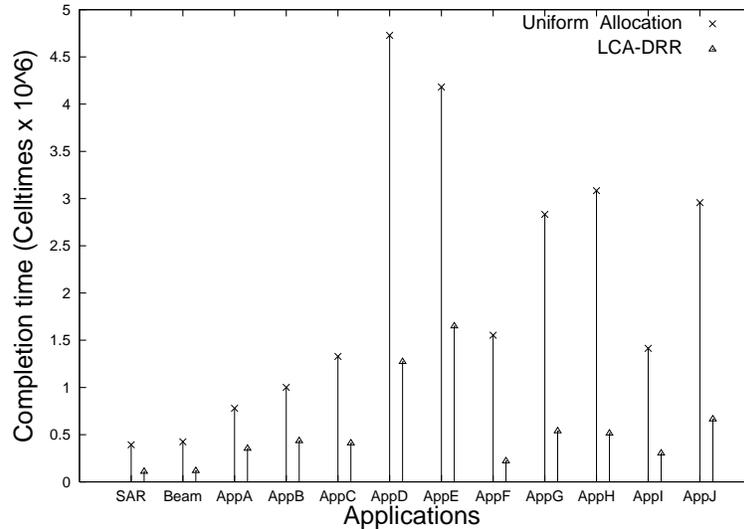
Figure 9: Communication phase completion times across application (with and without reconfiguration).

problems that are commonly encountered in embedded systems. Speedups of 1.9 to 7.1 were reported for the communications phases of the applications, corresponding to overall performance gains ranging from 20% to over 230%.

The reconfiguration exploited in this paper relies upon *a priori* knowledge of the bandwidth demands of the application. This is reasonable for the embedded application class described here, but in general one may not have such information available. We are currently investigating the performance of the optical interconnect in the context of an unknown (and unpredictable) communications workload, including the development of appropriate reconfiguration control algorithms as well as performance assessment of the system.

References

[1] R. Chamberlain, Ch'ng Shi Baw, M. Franklin, C. Hackmann, P. Krishnamurthy, A. Mahajan, and M. Wrighton. Evaluating the performance of photonic interconnection networks. In *Proc. of the 35th Annual Simulation Symposium*, April 2002.

[2] R.D. Chamberlain, M.A. Franklin, and A. Mahajan. VLSI photonic ring interconnect for embedded multicomputers: Architecture and performance. In *Proc. of 14th Conf. on Parallel and Distributed Computing Systems*, August 2001.

[3] 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.

[4] 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.

[5] L. Coldren, E. Hegblom, Y. Akulova, J. Ko, E. Strzelecka, and S. Hu. Vertical-cavity lasers for parallel optical interconnects. In *Proc. of 5th Int'l Conf. on Massively Parallel Processing Using Optical Interconnections*, pages 2–10, June 1998.
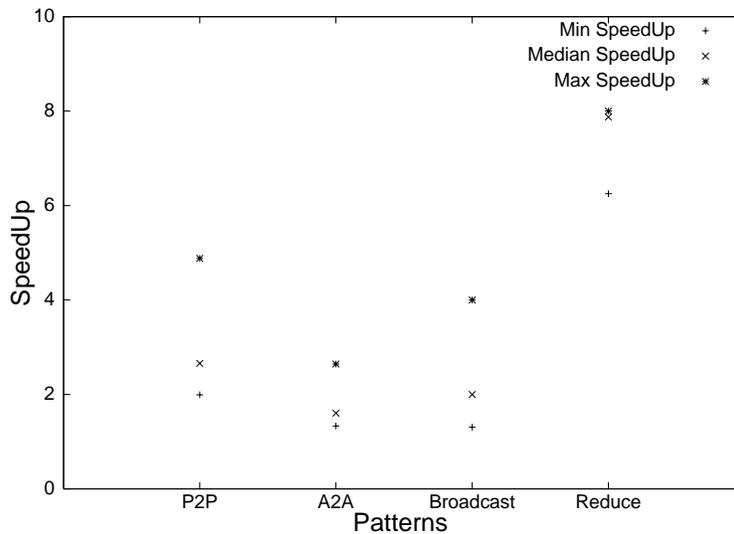
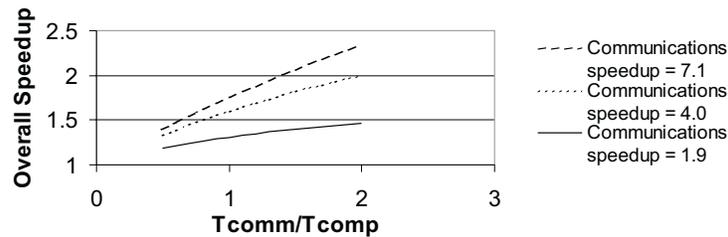Figure 10: Maximum, median and minimum speedup obtained across comm. patterns.

Figure 11: Overall performance improvement.

[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. Special issue on optical interconnections for digital systems. *Proceedings of the IEEE*, 88(6), June 2000.

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

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

[10] D. Plant, M.B. Venditti, E. Laprise, J. Faucher, K. Razavi, M. Chateauneuf, A.G. Kirk, and J.S. Ahearn. 256-channel bidirectional optical interconnect using vcsels and photodiodes on CMOS. *IEEE/OSA Journal of Lightwave Tech.*, 19(8):1093–1103, 2001.

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